Individualised evoked response detection based on the spectral noise colour

Jaime A. Undurraga [1, 4], **Michael Chesnaye** [2, 3], **David Simpson** [2], **and Søren Laugesen** [1]

[1]Interacoustics Research Unit, Technical University of Denmark, Ørsteds Plads, Building 352, DK-2800 Kgs. Lyngby, Denmark

[2]Institute of Sound and Vibration Research, Faculty of Engineering and the Environment, University of Southampton, Southampton, UK

[3]National Acoustic Laboratories, 16 University Avenue, Australian Hearing Hub, NSW 2109, Australia

[4]Department of Linguistics, 16 University Avenue, Macquarie University, NSW 2109, Australia

Lead Contact: *jmuu@iru.interacustics.com (J.U.)*

Correspondence: *jmuu@iru.interacustics.com (J.U.)*

Objective detection of evoked potentials (EPs) is central to digital diagnostics in hearing assessment and clinical neurophysiology, yet current approaches remain time-intensive and sensitive to inter-individual noise variability. Many existing detection methods rely on population-based assumptions or computationally demanding procedures, limiting robustness and efficiency in real-world clinical settings. We present Fmpi, a digital EP detection framework enabling individualised, real-time response detection through analytical modelling of the spectral colour and temporal dynamics of background noise within each recording. Using extensive simulations and large-scale human electroencephalography datasets spanning brainstem, steady-state, and cortical EPs recorded in adults and infants, we demonstrate performance comparable or superior to state-of-the-art bootstrapped methods while operating at a fraction of the computational cost and maintaining well-controlled sensitivity with improved specificity. Importantly, Fmpi incorporates a futility detection mechanism enabling early termination of uninformative recordings, reducing testing time without compromising diagnostic reliability.

evoked potentials | signal detection | objective detection | spectral colour | ABR | ASSR | Cortical |

INTRODUCTION

When we hear a sound, look at objects, move, or touch something, millions of neurons are activated, generating electromagnetic signals at different stages throughout the nervous system. These electromagnetic signals are known as evoked potentials (EPs) and play a crucial role in clinical diagnosis, intraoperative testing and neuroscience research. EPs can be detected non-invasively using extracorporeal measurements, such as magneto- and electro-encephalography (MEG and EEG), or invasively by placing sensors in direct contact with the neural source (e.g., electrocorticography).

Clinically, EPs are used to assess and diagnose sensory-neural dysfunction along the neural pathway for different sensory modalities. For example, auditory brainstem responses (ABRs)—an early auditory evoked potential (AEP) reflecting hearing function from the auditory nerve to the midbrain—play a critical role in newborn hearing screening[1], the diagnosis of auditory neuropathy[2] (a dysfunction in the transmission of action potentials along the auditory nerve), and intraoperative monitoring of auditory function[3]. Similarly, cortical AEPs—reflecting neural activity evoked by sounds at the cortex—play an important role for clinicians and primary caregivers in determining strategies to manage hearing loss[4].

Beyond the auditory domain, visual evoked potentials (VEPs), somatosensory evoked potentials (SEPs), and vestibular evoked myogenic potentials (VEMPs) are used clinically for the diagnosis of multiple sclerosis, intraoperative monitoring, diagnosis of optic nerve tumors or optic neuropathy, and assessment of vestibular function, among other applications[5].

Regardless of the sensory modality being tested, the measurement of an EP is, in principle, relatively straightforward. One or several types of stimuli are presented repeatedly—ranging from a few hundred repetitions (e.g., for cortical EPs) to several thousand (e.g., for brainstem EPs). The data are then filtered to reduce noise outside the EP bandwidth. Finally, data blocks aligned to each stimulus onset (epochs or trials) are averaged to further reduce noise within the bandwidth of interest.

The stimulus presentation rate—or, more generally, the rate of the specific cue within the stimulus that evokes the EP response—plays a crucial role in determining the type of neural response elicited. For example, low presentation rates—relative to the time course of the response—are usually employed when the morphology of the EP is of interest.

Low presentation rates evoke a transient-like EP, where the relatively long interval between presentations prevents overlap of consecutive EPs (e.g., ABRs[6], auditory change complex (ACC) responses[7], VEPs[8–10], VEMPs[11]) and minimise reduced neural responsiveness due to prior excitation (rate adaptation). When the stimulus cue evoking the EP is presented periodically and at higher rates, the morphology of the EP will be determined by the extent of overlap between consecutive responses, and it will follow—entrain to—such periodicity, generating the so-called steady-state EP (e.g., auditory steady-state responses (ASSRs)[12,13], interaural phase modulation following responses (IPM-FRs) and interaural time difference following responses (ITD-FRs)[14,15], auditory change following responses (AC-FRs)[16], steady-state VEPs[17,18], somatosensory steady-state responses[19], or steady-state VEMPs[20]).

Despite these relatively straightforward approaches to eliciting conventional EPs, accurately and efficiently determining their presence or absence remains a challenging topic. Decades of research have led to the development of multiple approaches to detect EPs in the time domain, frequency domain, or both domains. These include statistical[21–44] and machine-learning approaches[45–53].

Whilst these approaches have been useful for research and clinical diagnosis, their full potential has not yet been realised, partly due to the assumptions and constraints inherent to each approach. For example, assumptions about the properties of the underlying background noise (BN) in the recordings are often violated in practice[54–56]. The need for feature selection (e.g., the specific number and type of features used by statistical or machine learning approaches) varies depending on the type of EPs, and sub-optimal choices may lead to a loss of information, prolonged detection times, and reduced correct EP detection rates—specificity. Additionally, the requirement for high computational power (e.g., bootstrapping methods) can hinder implementation in existing clinical devices. Often, a combination of these factors limits the effectiveness of current approaches.

To ensure sufficient test sensitivity—here defined clinically as the correct identification of the absence of an EP—certain detection criteria rely on the population-level distribution of a given test statistic[24,25,29,37,38,57] rather than the specific statistical characteristics of an individual recording. Other approaches estimate EP detection thresholds from features extracted directly from the recording, albeit informed by fixed population-derived features[36,40,41,43,53]. Whilst these strategies typically ensure a minimum sensitivity of 95% in clinical contexts, the lack or limited degree of individualisation often results in elevated detection thresholds and unnecessarily prolonged test durations.

Improvements in the estimation of response detection thresholds for each individual recording can be expected to lead to more accurate and efficient diagnoses. This is particularly important in the context of hearing, as health systems are increasingly under pressure due to the rising prevalence of hearing loss worldwide[1,58]. Reducing testing time whilst maintaining well-controlled sensitivity, preferably using existing clinical hardware, is therefore crucial to alleviate the burden on already overwhelmed health centres.

Here, we investigate and validate a new general-purpose EP detection method, termed *Fmpi*. Fmpi is a parametric method based on the well-established F-test[24,25,29] which leverages information from within and across epochs to estimate the signal-to-noise ratio (SNR) and its statistical significance. However, rather than relying on assumed or empirical population estimates of underlying distribution parameters, Fmpi exploits the temporal and spectral properties (i.e., the 'colour') of the background noise to provide an *individualised* response detection threshold. Using both simulations and AEPs from a large body of research and clinical human data—including peripheral, cortical, and steady-state EPs from adults and infants with and without hearing loss—we demonstrate that Fmpi can detect EPs in both the time and frequency domains, in real time, with near-nominal sensitivity and improved specificity. Its performance matches that of computationally demanding bootstrapping methods and outperforms other state-of-the-art parametric methods. Furthermore, we develop and validate a novel Fmpi-based futility detector, which enables early termination of testing when strong evidence indicates that a response is unlikely within the maximum permissible testing time. This feature significantly reduces unnecessary testing time.

Overall, Fmpi offers significant benefits for both research and clinical diagnosis. It enables individualised and efficient detection, controlled sensitivity, low computational cost, and reduced testing time, while providing a more tangible—SNR based—interpretation than some other detection methods, making it ready for implementation in existing research and diagnostic devices.

RESULTS

We implemented three versions of the Fmpi algorithm, each differing in their approach to handling temporally non-stationary noise through distinct temporal weighting strategies (see Methods). These strategies include: (1) standard average (Fmpi-SA), where all epochs are weighted equally; (2) weighted average within epoch (Fmpi-WWE), where each epoch is weighted based on its internal variance; and (3) weighted average across epochs (Fmpi-WAE), where subsets of consecutive epochs share the same weights. We compared these Fmpi variants with several established parametric algorithms, including the clinical established time-based weighted average method (Fmp-WAE[24,25,29]), a time-based Hotelling-T2 (HT2[36,40,41,43])—a multivariate generalisation of Student's t-statistic—, two frequency-based Hotelling-T2 (HT2-f and HT2-f-diag)[44], and two frequency-based, fully bootstrapped methods: the non-parametric Q-Sample test[33,34,37,38,43] and

the modified T2 statistics test (T2-f-diag[44]).
The primary distinction between the Fmpi and Hotelling-T2 variants lies in the estimation of the degrees of freedom (DOF). The DOF (v_1 and v_2) are the two parameters that determine the EP response detection threshold from the Fisher–Snedecor probability distribution—F-distribution—assumed by both methods. The Fmpi estimates both DOF based on the individual spectral characteristics—i.e., the 'colour'—of the background noise and its non-stationary temporal fluctuations (see Methods). In contrast, Hotelling-T2 based methods, although capable of accounting for individual temporal correlations in the data, rely on fixed DOF determined by the dimensionality of the selected feature set (e.g., mean amplitudes from multiple time windows or frequency components) and the number of epochs.
The Fmp-WAE method, which also assumes an F-distribution, sets one DOF (v_1) to a fixed value, derived empirically from human EEG recordings[24,25,29]. The Q-Sample method transforms frequency-based features into ranks to reduce the effects of large non-stationary noise events. The response detection threshold is set according to the empirical properties of the underlying noise, estimated using bootstrapping[59,54,56]. The T2-f-diag statistic, derived from the Hotelling-T2, assumes that frequency-based features are uncorrelated to estimate the test statistic[44]. Like the Q-Sample, the T2-f-diag sets the response threshold using bootstrapping, which is computationally intensive but highly accurate.

Simulation of EPs in stationary and non-stationary noise

To evaluate detector performance, we conducted realistic simulations of cortical responses (Figure 1a) embedded in $1/f$ power spectral density (PSD) noise, which approximates brain background activity[60]. Simulations were performed under both stationary and non-stationary noise conditions, the latter incorporating eye-blink artefacts while maintaining all other parameters identical to the stationary condition. Parallel simulations without the target cortical signal ('no-response' data) were used to quantify detector sensitivity. First, we characterised the no-response data by examining the distribution of probability values (p-values) for each detector's test statistic (the F ratio for parametric tests, Q- and T2-values for bootstrapped tests; see Methods). Ideally, these p-values should be uniformly distributed[61], indicating unbiased detection. Deviations from uniformity reflect bias; the larger the deviation, the greater the bias (Figure 1b). We also quantified false response detection rates as a function of the number of epochs, using a nominal sensitivity of 95% (Figure 1c).

In stationary noise simulations without a response, all detectors except Fmp-WAE and HT2-f-diag produced nearly unbiased estimates, as indicated by the flat p-value distributions (Figure 1b, stationary) and close agreement between empirical (black lines) and theoretical (orange lines) probability density functions (PDFs) of the F-distribution (insets in Figure 1b, stationary). In non-stationary noise, Fmp-WAE, Fmpi variants, and HT2 methods exhibited some bias (Figure 1b, non-stationary), which was also reflected in false response detection rates (Figure 1c). Under stationary noise (Figure 1c, left panel), Fmpi variants, HT2, HT2-f, Q-Sample, and T2-f-diag methods achieved nearly ideal false response detection rates of 5% (95% nominal sensitivity), both within and across all epochs (Figure 1c, left inset). The Fmp-WAE and HT2-f-diag detected fewer false responses than expected, indicating a bias towards higher sensitivity. In non-stationary noise, the Fmpi-WWE also showed a lower false response detection rate than the nominal value (Figure 1c, right panel).
We then assessed true response detection rates by adding the simulated cortical responses to the noise. In stationary noise, final detection rates were: Fmpi variants (94.5-96.2%), bootstrapped Q-Sample (96.3%), Hotelling-T2 variants (87.2-95.8%), Fmp-WAE (71.8%), and bootstrapped T2-f-diag (96.0%) (Figure 1d, left). Under non-stationary noise, overall detection rates decreased (Figure 1d, right). The Fmpi-WWE outperformed other methods with a final detection rate of 54.9%, followed by the bootstrapped Q-Sample (48.4%). These detection rates were nearly double those of the remaining detectors, which ranged from 11.7% (Fmp-WAE) to 28.8% (HT2-f). Notably, the Fmpi-WWE showed higher detection rates despite a bias towards greater sensitivity in the absence of a response.

[Figure 1 about here.]

To further evaluate detector performance across different sensitivity values, we calculated several metrics. First, we compared nominal sensitivity (Figure 2a, solid black lines) to observed sensitivity (Figure 2a, coloured lines). Consistent with the initial results, observed sensitivity for all methods except Fmp-WAE and HT2-f-diag closely matched the nominal sensitivity under stationary noise (Figure 2a, left). The Fmp-WAE showed a better fit in non-stationary noise (Figure 2a, right panel).
Focusing on the clinically relevant sensitivity range (95-100%; Figure 2a, insets), the Fmp-WAE consistently showed a bias towards higher sensitivity in both stationary and non-stationary noise. The average bias (mean absolute difference between nominal and actual sensitivity) for stationary vs. non-stationary noise was: Fmp-WAE (2.57 vs. 1.33%), Fmpi-SA (-0.18 vs. 0.11%), Fmpi-WWE (0.05 vs. 1.63%), Fmpi-WAE (-0.28 vs. -0.12%), Q-Sample (-0.08 vs. 0.17%), HT2 (-0.19 vs. -0.98), HT2-f (-0.49 vs. -0.73%), HT2-f-diag (0.80 vs. 0.70 %), and T2-f-diag (-0.15 vs. 0.05 %), with the sign indicating a higher—positive—or lower—negative—average bias relative to the nominal value.

We then assessed performance as a function of the number of epochs (i.e., time) by generating sensitivity-specificity

curves—receiver operating characteristic curves (ROCs)—for the full sensitivity range (Figure 2b) and the clinically relevant range (Figure 2c). The highest performance over time—indicated by ROCs shifted towards the upper left corner—was observed for all Fmpi variants, bootstrapped Q-Sample, bootstrapped T2-f-diag, and frequency-based HT2 methods under stationary noise, and for Fmpi-WWE and bootstrapped Q-Sample under non-stationary noise. This pattern held across all sensitivity levels (Figure 2b) and within the clinically relevant range of the ROCs (Figure 2c), as shown by the area under the curve (AUC) of the clinically relevant ROC (normalised by the total area within the relevant sensitivity region) over time (Figure 2d; see Methods).

We summarised overall performance by calculating the average area under the curve (AAUC) of the clinically relevant AUCs (Figure 2d) across all epochs (Figure 2e). AAUC values from 1000 simulations (treated as data from 50 participants) were analysed using a linear mixed-effects (LME) model, with AAUC as the dependent variable and detector type as the independent variable. The analysis revealed a significant effect of detector type [$p < 0.001$]. Post hoc comparisons confirmed that the Fmpi-SA, Fmpi-WAE, Q-Sample, and T2-f-diag significantly outperformed other detectors in stationary noise, whereas the Fmpi-WWE was superior under non-stationary noise (as indicated by non-overlapping error bars in Figure 2e).
To determine whether performance improvements were attributable to differences in observed sensitivity rather than intrinsic statistical properties of the detectors, thresholds for each detector were adjusted such that observed and nominal sensitivities were equal within the clinically relevant range. Following this equalisation the AUC remained largely unchanged under stationary noise, with improvements observed only for the Fmp-WAE and HT2-f-diag. This outcome was anticipated, as the Fmp-WAE and HT2-f-diag measure the same statistical quantities as the Fmpi-WAE and T2-f-diag, respectively. Under non-stationary noise, equalisation yielded results similar to the original analysis, with the notable exception of Fmpi-WWE, whose final AUC improved from 49.5 (unequalised) to 53.4% (equalised).

The generalisability of the Fmpi to other EP types was confirmed via simulations of ABRs (Sup. Figure 1) and ASSRs (Sup. Figure 2). The results indicated that Fmpi variants and Q-Sample approaches consistently yielded similar or superior performance relative to other detectors.

[Figure 2 about here.]

Having demonstrated the reliability of the Fmpi as a detector, its capacity for futility detection was systematically tested (see Methods). Specifically, we quantified the instances in which the estimated number of epochs to detect a cortical response, shown as a function of the actual number of epochs (Figure 3; each trace corresponds to an individual simulation), exceeded a predefined futility threshold of 200 epochs (Figure 3, horizontal line). In the presence of a cortical response under stationary noise (Figure 3, top-left panel), the trajectory of the estimated necessary epochs vs. the actual number of epochs typically crossed below the unitary diagonal—the detection boundary (Figure 3, diagonal black solid line)—prior to reaching the futility threshold, indicating that, in most simulations, the number of averaged epochs was sufficient to yield a significant response before futility was reached (Figure 3, green traces). Conversely, under non-stationary noise conditions (Figure 3, top-right panel), the futility threshold was surpassed at a substantially higher rate. That finding is consistent with the observed response detection rates (Figure 1d), wherein $\approx 94.5\%$ of responses were detected at 200 epochs under stationary noise, compared to only $\approx 54.9\%$ detected by Fmpi-WWE under non-stationary noise.

To further assess the efficacy of the futility detector, we calculated the proportion of 'expected' futile measurements, defined as the ratio between the number of no detected responses at the end of the measurement period (at a nominal sensitivity of 95%) and the total number of measurements. For cortical responses in stationary noise, the expected futility was 5.5% whilst the observed futility rate—defined as the proportion of responses crossing the futility threshold at any time—was 7.8%, yielding a difference of 2.3% between predicted and observed futility. In non-stationary noise, the expected and observed futility rates were 45.1 and 53.0%, respectively. In the absence of a simulated cortical response, the expected and observed futility rates were 95.4 and 96.1% in stationary noise, and 97.9 and 99.0% in non-stationary noise, respectively. Absolute agreement between observed and predicted futility across all data, i.e., stationary and non-stationary noise with and without cortical responses, was 97% and was significant, as confirmed by Cohen's Kappa coefficient (κ)[62] of 0.94 [$p < 0.001$].
Practically, these results imply that, when cortical responses are absent, half of the recordings would have terminated after 140 epochs (the median) in stationary noise and after 120 epochs in non-stationary noise, corresponding to 30% and 40% time savings, respectively, while maintaining error rates at the specified test sensitivity. Comparable results were obtained in simulations of ABRs and ASSRs (data not shown).

[Figure 3 about here.]

Detection of EPs in human recordings

Having shown that the Fmpi and bootstrapped Q-Sample performed comparably, and frequently outperformed other detectors in simulated data, we subsequently evaluated their performance using EEG recordings from human participants. Specifically, EPs elicited by auditory stimuli from human participants across ten different studies (Table 1). EPs included ABRs, ASSRs, and cortical responses, with data collected from both adults and infants, including individuals with and without hearing loss, in a variety of clinical and research contexts. While the specific experimental conditions are detailed in Table 1, for the purposes of this analysis, only conditions in which EPs were expected were selected.
For time-domain analyses (i.e., ABRs and cortical responses), no-response datasets—required for estimating sensitivity, false response detection rates, and AAUCs—were generated by removing the EP from the original recordings and substituting the original event times with randomised events (see Methods). For frequency-domain analyses (i.e., ASSRs), the original event times were replaced with randomised ones and no-response analyses were carried out on neighbouring frequency bins adjacent to the target frequency, where no response was anticipated.

[Table 1 about here.]

First, we analysed cortical EPs evoked by changes in ongoing stimuli (studies 1-3, ACC responses from adults) and by speech-like sounds (study 4; onset cortical responses from infants using hearing aids). The magnitude of the responses, which was influenced by stimulus characteristics, participants' hearing status and age, as well as the total number of epochs, varied across studies (Figure 4a).

Consistent with simulation results, the greatest sensitivity bias was observed for the Fmp-WAE and HT2-f-diag detectors (Figure 4b). In contrast, all other methods exhibited a small bias, averaging less than 1% (Figure 4b, insets). This trend was also reflected in the false response detection rates at a nominal sensitivity of 95% (Figure 4c). Notably, both the response detection rates at 95% nominal sensitivity (Figure 4d) and the AAUC (Figure 4e)—which summarises performance across epochs for all clinically relevant sensitivities—were highest for the Fmpi-WWE and Q-Sample methods.
This finding was corroborated by an LME model applied independently to each study, with AAUC as the dependent variable, *detector* as the fixed effect, and *subject* as a random effect. In all studies, the *detector* factor was significant [$p < 0.001$]. Post hoc comparisons, adjusted for family-wise error rate (FWER) using the Holm method, indicated that the Fmpi-WWE and Q-Sample detectors consistently achieved the highest AAUC across all studies. No significant differences were observed between the Fmpi-WWE and Q-Sample methods in any study (as indicated by overlapping error bars in Figure 4e), and both methods performed significantly better than, or comparable to, the best of the remaining methods. The Fmp-WAE approach consistently exhibited the lowest performance, whilst the performance of the various HT2 methods and bootstrapped T2-f-diag varied by study.
In study 1, the Fmpi-WWE and Q-Sample outperformed the Fmp-WAE, HT2 and HT2-f-diag [$p < 0.05$]. In study 2, the Fmpi-WWE and Q-Sample outperformed all HT2 variants and bootstrapped T2-f-diag [$p < 0.05$]. In study 3, all Fmpi variants, bootstrapped Q-Sample, and T2-f-diag performed similarly [$p > 0.05$], although the Q-Sample method outperformed all HT2 variants. In study 4, only the Fmp-WAE detector demonstrated poor performance [$p < 0.05$]. Notably, results from study 2 paralleled findings from non-stationary noise simulations, demonstrating improved performance for the Fmpi-WWE and Q-Sample (cf. Figure 4d and Figure 1d).
These results confirm that the detection of cortical EPs—characterised by longer and fewer epochs—is most effectively achieved using the Fmpi-WWE, with performance equivalent to the more computationally intensive bootstrapped methods.

[Figure 4 about here.]

Subsequently, the detection methods were applied to ABRs (Figure 5a) obtained from both adults (studies 5 and 7) and infant participants (study 6). This dataset represented the largest and most computationally demanding cohort analysed, comprising ≈ 3000 independent measurements. The computational burden was primarily attributable to the bootstrapped methods (Q-Sample & T2-f-diag) , which required substantial processing time; for instance, processing a single block of 4000 epochs necessitated over 50 seconds using six parallel processes, whereas parametric detectors (Fmp-WAE, Fmpi, and HT2 variants) completed the same task in a fraction of a second on a single processor.
The observed bias (Figure 5b) and false response detection rates (Figure 5c) were analogous to those identified in the cortical response analyses. However, the HT2-f-diag approach exhibited improved performance (small degree of bias) compared to the Fmp-WAE method. The influence of the *detector* factor on AAUC was substantiated by an LME model [$p < 0.001$ across all three studies]. In accordance with simulation results, the Fmpi-WWE and Fmpi-WAE demonstrated superior performance among all Fmpi variants (Figure 5d&e). The performance of the Fmpi-WWE and Fmpi-WAE did not differ significantly from that of the Q-Sample method [$p > 0.05$ in all three studies, as indicated by overlapping error bars in Figure 5e]. Both Fmpi-WWE and Q-Sample methods significantly outperformed Fmpi-SA, Fmp-WAE, HT2, HT2-f, HT2-f-diag, and T2-f-diag [$p < 0.05$] in studies 5 and 6, whereas, in study 7, only the Fmp-WAE exhibited poor performance.

These results confirm that ABR detection is most effectively achieved using weighted averaging approaches (Fmpi-WAE and Fmpi-WWE).

[Figure 5 about here.]

Following the demonstration of the Fmpi's robust performance for time-domain EPs, its efficacy in detecting ASSRs in the frequency domain was evaluated. Analyses encompassed ASSRs at high modulation rates (39 to 81 Hz; Figure 6), which are commonly employed in clinical practice, as well as at low modulation rates (≈ 7 Hz; Figure 7). The dataset included normal hearing (NH) adults (studies 8 and 9) and hearing-impaired (HI) infants using hearing aids (study 10). All detectors were provided with identical frequency sets: the modulation rate and the next two harmonics (studies 8 and 9), and the next eleven harmonics (study 10; see Methods).

Consistent with time-domain analyses, the greatest sensitivity bias was observed for the Fmp-WAE and HT2-f-diag detectors (Figure 6b&c and Figure 7b&c). This effect was particularly pronounced in study 10, where the Fmp-WAE—which assumes that each frequency contributes with equal DOF (see Methods)—resulted in reduced sensitivity, increased false response detection rates, and inflated detection rates (Figure 6b-d, study 10). Irrespective of modulation rate, the results corroborated simulation findings (Sup. Figure 2) with Fmpi variants and bootstrapped Q-Sample demonstrating superior performance (Figure 6d&e and Figure 7d&e).
LME modelling confirmed a significant effect of *detector* type on AAUC [$p < 0.001$ in all studies]. For high modulation rates ASSRs (Figure 6), post hoc comparisons revealed no significant differences among Fmpi-WWE, Fmpi-WAE, and Q-Sample (as indicated by overlapping error bars in Figure 6e). In all studies, the Fmpi-WAE, Fmpi-WWE, and bootstrapped Q-Sample outperformed HT2-f and HT2-f-diag [$p < 0.05$]. Additionally, in all but one study (study 8b), Q-Sample outperformed the bootstrapped T2-f-diag [$p < 0.05$], whilst Fmpi-WWE outperformed bootstrapped T2-f-diag in studies 8a and 9a [$p < 0.05$].

[Figure 6 about here.]

The results for low modulation rate ASSRs were analogous, with no significant differences between Fmpi variants and bootstrapped Q-Sample methods [$p > 0.05$; Figure 7a-e]. Both methods outperformed, or at least matched, the best performance of the remaining approaches.
Collectively, these findings indicate that the Fmpi is a reliable, general-purpose EP detector, characterised by well-controlled sensitivity and consistently high performance as reflected by AAUC. Importantly, this approach imposes minimal computational demands on existing clinical systems while maintaining performance comparable to bootstrapped methods.

[Figure 7 about here.]

Testing time and futility detection in human recordings

To further assess the clinical utility of the Fmpi detector, overall testing time—defined in terms of the number of epochs—was quantified for each study and detection method. Testing times were estimated from the response detection rates derived for each study, calculated as a function of the number of epochs for nominal sensitivities ranging from 95% to 100% in increments of 0.01%. These response detection rate functions, which describe the proportion of detected responses as a function of accumulated epochs at a given sensitivity level, were used to estimate the number of responses detected at each epoch from an initial pool of 10000 measurements. Once a response was detected, the corresponding measurement was removed from the pool, and no further testing was assumed. For each nominal sensitivity, testing times for all detectors were compared relative to the fastest detector, and the difference was normalised and expressed as a percentage of the shortest testing time. Positive values thus indicate additional testing time incurred relative to the fastest detector at a given sensitivity level. The mean additional testing time across all sensitivity levels was subsequently calculated to provide an overall estimate of average performance. To ensure fairness and comparability across detectors, reductions in testing time attributable to futility detection—which is only available for the Fmpi detector—were excluded from this analysis.

Consistent with their superior response detection rates, both Fmpi and bootstrapped Q-Sample detectors exhibited the shortest average testing times across all studies (Figure 8a; Table 2). For cortical responses, the average additional testing times were as follows: Fmpi-SA (8.8%), Fmpi-WWE (4.9%), Fmpi-WAE (6.8%), Q-Sample (0.3%), and T2-f-diag (7.1%), all of which were substantially shorter than those for Fmp-WAE (63.0%), HT2 (20.9%), HT2-f (13.8%), and HT2-f-diag (33.4%). For ABRs, Fmpi-WAE (3.3%), Fmpi-WWE (2.6%) and Q-Sample (1.1%) outperformed Fmp-WAE (44.2%; the clinical standard), Fmpi-SA (7.5%), HT2 (23.2%), HT2-f (8.8%), HT2-f-diag (11.2%), and T2-f-diag (7.2%). In the context of ASSRs, Fmpi-SA (5.8%), Fmpi-WWE (3.7%), Fmpi-WAE (4.6%), and Q-Sample (10.3%) again demonstrated superior efficiency compared to Fmp-WAE (27.7%), HT2-f (73.5%), HT2-f-diag (114.5%), and T2-f-diag (34.2%). These findings indicate that the Fmpi detector can substantially reduce testing time, achieving efficiency comparable to the computationally intensive bootstrapped methods, without compromising sensitivity or performance.

In addition to improved response detection time, the Fmpi offers further time savings through futility detection. This aspect was also investigated from EP recordings. Employing the same methodology as in the simulations, expected and predicted futility rates were compared based on a predefined futility threshold, defined as the total number of epochs available for each recording. Analyses were conducted for both the original data (Signal + Noise) and the no-response data derived from the same datasets (Noise). The futility curves derived from human data (Figure 8b; from study 5) closely paralleled those obtained in simulations (Figure 3).

The concordance between expected and predicted futility was highly significant (Figure 8c). Absolute agreement percentages (Table 3), reflecting the alignment between final detection and futility detection for each measurement, ranged from 87.5% to 100% across individual studies, with an overall agreement of 97.1% when aggregated across all studies. These findings were further supported by Cohen's Kappa coefficient analyses, with κ values spanning 0.87 to 0.97 across studies [$p < 0.001$ for all cases], and a pooled κ value of 0.94 [$p < 0.001$]. Notably, κ values exceeding > 0.80 are generally interpreted as 'almost perfect' agreement[63].

The proportion of time saved—defined as the ratio between the epoch at which futility was detected and the predefined futility threshold—varied across studies and datasets (Figure 8d). For the original recordings, time savings were inversely related to response detection rates, as expected. For instance, time savings of approximately 25% were observed for ASSRs in study 10 which exhibited a response detection rate of 43%, whereas savings were reduced to 13% for ASSRs in study 9b (80% detection rate) and to 9.8% for ABRs in study 5 (73% detection rate). In contrast, no time savings were observed for the 40 Hz ASSRs in study 8b, which demonstrated a response detection rate of 100%.

When the analysis was applied to no-response data, reflecting the maximal potential benefit of futility detection, time savings ranged from 23% for cortical recordings to 47% for ASSR data.

[Figure 8 about here.]

[Table 2 about here.]

[Table 3 about here.]

The results demonstrate that the Fmpi detector significantly shortens the time required to detect the presence of an EP or to determine when continued recording is unlikely to yield a result, particularly in scenarios where an EP must be obtained within a predefined time frame. Moreover, the detector is computationally efficient and compatible with existing clinical devices, offering the potential for meaningful improvements in clinical practice.

Discussion

In this study, we introduced and validated a novel method for detecting EPs—the Fmpi. By leveraging the unique spectral and temporal characteristics of background noise in EP recordings, the Fmpi enables accurate estimation of the DOF (v_y and v_{rn}) that define the F-distribution used to determine the response detection threshold. Through extensive simulations and analyses of human EP datasets, we demonstrated that the Fmpi achieves sensitivity and specificity comparable to those of computationally intensive bootstrapped methods, while outperforming other state-of-the-art parametric techniques. This performance was consistent across both time- and frequency-domain analyses, establishing the Fmpi as a general-purpose EP detector.

In addition to its statistical robustness, the Fmpi significantly reduced testing time by either providing faster EP detection or identifying futile recordings early during a recording. Under the analysis parameters employed, average time savings due to faster detection ranged from 6% to 111% across brainstem, midbrain, and cortical EPs, relative to other parametric methods, and were comparable to the bootstrapped Q-Sample. Notably, the Fmpi reduced detection time by 42% compared to the clinical standard (Fmp-WAE) for ABRs.

Furthermore, futility detection based on no response data derived from human recordings yielded additional time savings ranging from 23% to 46%, underscoring the method's potential to enhance clinical efficiency.

Fewer assumptions and more individualised detection

Traditional parametric detection methods, such as Fmp-WAE variants, rely on the F-distribution with empirically determined DOF (v_1 and v_2), which are typically tuned to specific experimental paradigms, recording configurations, and the statistical properties of the data used to parameterise the underlying distributions. These include, but are not limited to, the sampling rate, bandpass filter settings, artefact rejection criteria, and intrinsic spectral characteristics of the EEG signal. In practice, such tuning is achieved either by adjusting the DOF (as in the Fmp-WAE) or by modifying number of features included in the analysis, as in the HT2, to better accommodate the characteristics of a particular response type within a given population.

These adjustments are necessary because the serial correlation among samples within each epoch—which directly influence the DOF in Fmp-WAE variants and the covariance structure in HT2 variants—are strongly dependent on the background noise characteristics, sampling rate, the presence of artefacts (e.g., myogenic activity), and the applied filter settings (see Sup. Figure 3a&b).

Rather than relying on fixed DOF values or response-specific feature counts, the Fmpi addresses this challenge by incorporating all samples as features and dynamically estimating the DOF for each measurement. This estimation is based on both the spectral characteristics—i.e., the 'colour'—and the non-stationary temporal dynamics of the background noise. This individualised approach allows the Fmpi to accommodate a wide range of sampling rates, filter configurations, and non-stationary noise conditions, achieving performance levels comparable only to those of data-driven bootstrapped methods. Indeed, the Fmpi's analytical DOF estimations align well with previously reported empirical values. For example, a prior study reported an empirical DOF v_1 of 160 for a 10 ms time window, using stationary white noise high-pass filtered at 100 Hz, sampled at 16 kHz, and averaged across 250 epochs[24]. Applying the Fmpi's analytical solution for $1/f^p$ noise under the same conditions yielded an identical value (see Sup. 1.10 and Sup. Figure 3b, green marker for $p = 0$). Similarly, for pink noise bandpass filtered between 200 and 1000 Hz, the same study reported an empirical value v_1 of 15.0. The Fmpi's analytical solution for ideal filters produced a value of 13.0 (Sup. Figure 3b, green marker for $p = 1$), and 15.0 when using steep, non-ideal filters (not shown).
Further validation is provided by the seminal work by Elberling and Don[24], who estimated fixed DOF v_1 values by measuring multiple realisations from non-stimulus recordings in adult human participants, yielding a range of 8 to 22 DOF. These were subsequently fitted to a Gaussian distribution, with a fixed DOF v_1 of 5 corresponding to the 5% percentile threshold. Employing the same approach, the 5th percentile value of DOF v_1 estimated from adult ABR recordings (studies 5 and 7) was v_1 = 6.07, with v_1 ranging from 3.6 to 36.8.
In a recent publication, Lightfoot et al.[57] estimated the 97.5% and 99% F-value percentiles as 2.2 and 2.8, respectively, from no-response data from infants (33-1500 Hz bandpass filtered; 10 ms time window). The estimates from infants' data (study 6) were 2.1 and 2.5, respectively, whilst the 5th percentile of DOF v_1 was 6.3, ranging from 3.8 to 22.8.
The wide range of v_1 values estimated by the Fmpi is comparable to those shown by Elberling and Don[24] and Lightfoot et al.[57], and underlines the importance of using individualised rather than fixed DOF.

Estimates of optimal number of features—specifically, voltage mean intervals—for ABRs using the HT2 method in the time domain (30-1500 Hz bandpass filtered; 15 ms time window) in simulated data have been reported to range from 16 (corresponding to an interval of 938 µs) to 40 (375 µs)[40,53]. Analytical calculations of the DOF v_1, interpreted here as the effective number of independent samples within the same bandwidth and analysis window, yielded values between 14.1 (1 ms, pink noise) and 44 (340 µs, white noise).
In a previous study[36], cortical responses analysed with the HT2 in the time domain, employed a voltage mean interval of 50 ms (resulting in nine features for data bandpass filtered between 0.1 and 30 Hz, using a 450 ms analysis window), whilst recent studies have employed intervals ranging between 33 to 50 ms[41,43,44,64,65]. Analytical solutions for v_1 using filter parameters and analysis windows similar to those described by Golding et al.[36], produced time intervals ranging from 153 ms (2.93 DOF, pink noise) to 17 ms (16.72 DOF, white noise).
The 5th percentile of DOF v_1 estimated from adult data in studies 1 to 3 was 10.8 (corresponding to 42.0 ms interval), which falls within the range reported in those investigations. Given that features are typically selected from time windows in which responses are expected to be strongest, these results further indicate that HT2 optimisations are likely driven by the reduction of non-zero off-diagonal elements in the sample covariance matrix used to compute the HT2 statistic—that is, by selecting intervals for which noise samples are largely statistically independent.
Collectively, these results demonstrate that the Fmpi method accurately accounts for previously published data and that its superior performance arises from its ability to adapt the DOF to the broad spectrum of values encountered in EP recordings.

Analysis parameters

In this study, we sought to harmonise the parameter space—including filter settings, analysis windows, and feature selection—across all detectors, aligning them as closely as possible with those utilised in prior investigations[40,44]. No attempt was made to optimise detector performance specifically for the datasets analysed herein. However, for comparative purposes, we carried out a secondary analysis (not shown) to ensure equivalent signal energy in both the time and frequency domains across all detection methods. This was accomplished by applying bandpass filtering to the time-domain data such that the bandwidth matched that of the frequency-domain detectors within the same analysis window, thereby guaranteeing that all detectors operated on data with comparable signal energy. The analysis revealed that frequency-based HT2 detectors exhibited the highest sensitivity bias. However, the overall pattern of results remained consistent, with the Fmpi and Q-Sample detectors demonstrating superior performance.
An additional critical consideration in time-domain analyses is the relationship between the length of the analysis window and the cut-off frequency of the high-pass filter[24,29]. Previous studies employing the Fmp-WAE method for cortical

and ABR analyses have often overlooked this relationship[40,43,53,57], potentially introducing biases. Specifically, the F-value—which quantifies the ratio between the variance of the average response within the analysis window and the variance of the residual noise (RN) across epochs—can be distorted by an imbalance in the energy content of these two quantities. For instance, if the high-pass filter is set at 30 Hz and the analysis window (T_w) is 10 ms, the variance of the RN will include energy from all frequencies ≥ 30 Hz, whereas the variance of the averaged signal, further high-pass filtered by removal of the mean (demeaning) within the analysis window, will mostly reflect frequencies ≥ 100 Hz. This mismatch results in biased F-values and diminished response detection rates, with the extent of bias being strongly influenced by the spectral content of the noise; greater low-frequency energy exacerbates the bias.

This issue is also pertinent to the Fmpi method and can be mitigated either by a high-pass filter with a cut-off frequency above $1/T_w$[24], or by using the same analysis window to estimate both variances, as is done in the HT2 approach for time-domain analyses.

These technical considerations should be further analysed for potential incorporation in future revisions of clinical guidelines (e.g. those of the British Society of Audiology).

Finally, in this study, we intentionally refrained from applying epoch-rejection procedures during the analyses. This enabled the evaluation of accuracy and robustness of all detection methods under conditions of non-stationary noise. Although implementing rejection procedures (data not shown) generally improved the performance of detectors that initially exhibited suboptimal results, the overall performance hierarchy remained similar. The most substantial improvements were observed in cortical recordings when applying a ±50 µV rejection criterion, whereas ASSR recordings (using a ±50 µV rejection criterion) and ABRs (excluding 10% of epochs with the greatest variance as rejection criterion), demonstrated minimal improvements.

Limitations and further directions

The Fmpi, like any other detection method, requires correction when sequential testing is employed. This adjustment, which was not applied to any detector in the present study when estimating testing time or futility, is necessary to mitigate the potential reduction in sensitivity that may arise from the increased probability of false-positive detections due to multiple comparisons[66]. Nevertheless, we anticipate minimal deviation from the overall findings, as the robust performance of the Fmpi was consistently maintained across clinically relevant sensitivity ranges.

In this study, we have also developed and validated a futility detection approach, which is based on the premise of well-established measurement limits—defined either by a maximum testing time or by a maximum number of epochs—for different EPs and stimuli. These limits should be determined according to the specificity observed in normative populations, i.e., a testing time that ensures the desired EP detection rates in a normal population. It is important to emphasise that futility detection operates on probabilistic principles. Consequently, when futility is detected, it should prompt a careful evaluation of the recording conditions, such as noise levels, sensor contact quality, and stimulus delivery fidelity. If suboptimal conditions are identified—such as instances where participants are unable to remain still, resulting in excessive myogenic noise—retesting may be warranted to ensure data quality.

Furthermore, both the Fmpi and the other detection methods evaluated in this work assume that the noise is independent across epochs, an assumption typically achieved through high-pass filtering of the data. However, this assumption may not hold when spatial filters are used to denoise the data (e.g., denoising source separation[67]), or when detection is performed by averaging data across different recording channels (such in multichannel EEG) rather than, or in addition to, across epochs. Further research should investigate whether the Fmpi and other parametric approaches can be extended to account for dependencies across both temporal and spatial dimensions.

Finally, although the analysis parameters employed here are broadly consistent with those used in previous studies, further methodological refinements warrant investigation. These include optimisation of analysis windows, filter bandwidths, and the number of frequency components used for response detection and noise estimation in frequency-domain analyses. Such refinements may further enhance detection performance and robustness across a wider range of experimental and clinical recording conditions.

METHODS

Fmpi

Most evoked potentials (EPs) are estimated by repeatedly presenting the evoking stimulus (Figure 9a), thereby reducing the underlying background noise (BN) contaminating each presentation, ($y_i(t)$), and preserving the deterministic components of the evoked response through averaging. Once the average response is obtained, it is critical to assess whether sufficient evidence exists to consider the averaged waveform a true EP, rather than a random realisation of the post-averaged residual noise (RN).

The Fmpi is a parametric test that, like previous methods[24,25,29], measures the ratio between the estimated variance of the averaged EP, $\hat{\sigma}_y^2$, and the estimated variance of the residual noise, $\hat{\sigma}_{\mathrm{RN}}^2$, as shown in Eq. 1

$$F = \frac{\hat{\sigma}_y^2}{\hat{\sigma}_{\mathrm{RN}}^2} \tag{1}$$

The probability that the F-value belongs to the null distribution, which represents the absence of a response, is given by Eq. 2

$$p = 1 - F_{CDF}\left(F, v_y, v_{\mathrm{RN}}\right) \tag{2}$$

where $F_{CDF}\left(F, v_y, v_{\mathrm{RN}}\right)$ is the cumulative distribution function of an F-distributed random variable with degrees of freedom (DOF) v_y and v_{RN}.
Both v_y and v_{RN} are critical, as these parameters are used to set the threshold for objectively determining the presence or absence of an EP. The DOF are determined by the characteristics of the background noise, specifically its spectral properties and degree of stationarity, both of which are expected to exhibit variability across individual recordings. The Fmpi provides individualised estimates of v_y and v_{RN} for EPs analysed using either the time or the frequency domains, based on the properties—specifically, the spectral 'colour'—of the electroencephography (EEG) power spectral density (PSD). A detailed description and derivation of these estimates is provided in the Supplementary information.
The determination of the variances and their corresponding DOF depends on the analysis domain (time or frequency), as explained below.

Time-domain EP

Time-domain EPs are typically measured by presenting repeated stimuli that evoke the response of interest (Figure 9a, purple trace). However, these recordings are contaminated by biological and endogenous noise (Figure 9a, grey trace). To mitigate this noise, the data are usually bandpass filtered to remove frequencies outside the EP bandwidth and segmented into epochs, $y_i\left(t\right)$, of constant duration T, covering the response time of interest. Noisy epochs may be excluded based on an empirical threshold or downweighted (Figure 9b) by applying per-epoch weights (w_i, with $\sum_i w_i = 1$). To further reduce noise within the EP bandwidth (Figure 9b&c, grey traces) and approximate the true response (Figure 9 purple lines), the weighted epochs are averaged (Eq. 3; Figure 9d):

$$\overline{y}\left(t\right) = \sum_{i=1}^{N} w_i y_i\left(t\right) \tag{3}$$

The post-average RN (see Supplementary information for details) can be estimated as:

$$\hat{\sigma}_{\mathrm{RN}}^2 = \sum_{i=1}^{N} w_i^2 \hat{\sigma}_{BN_i}^2 \tag{4}$$

where the vector of weights is defined as:

$$\boldsymbol{w}_w^{\top} = \frac{\left[\frac{1}{\hat{\sigma}_{\mathrm{BN}_1}^2}, \ldots, \frac{1}{\hat{\sigma}_{\mathrm{BN}_N}^2}\right]}{\left(\sum_{i=1}^{N} \frac{1}{\hat{\sigma}_{\mathrm{BN}_i}^2}\right)} \tag{5}$$

with $\hat{\sigma}_{\mathrm{BN}_i}^2$ representing the per-epoch BN variance (described later) and with $()^{\top}$ denoting the transpose operator.
For time-domain EPs, it can be shown (Supplementary information) that the DOF of the RN variance (v_{RN}), estimated from all available samples within each epoch of duration T are given by:

$$v_{\mathrm{RN}} = v\frac{\left[\sum_{i=1}^{N} w_i\right]^2}{\sum_{i=1}^{N} w_i^2} = vN_e \tag{6}$$

where N_e is the effective number of epochs estimated from the weights ($N_e \leq N$) and v are the DOF within each epoch estimated from Eq. 7.

$$v = \frac{2\left[\sum_{j=1}^{N_\omega} \hat{G}_{\eta\eta}\left(\omega_j, T\right)\right]^2}{\sum_{j=1}^{N_\omega}\left[\hat{G}_{\eta\eta}\left(\omega_j, T\right)\right]^2} \tag{7}$$

where $\hat{G}_{\eta\eta}\left(\omega_j, T\right)$ is the PSD (Figure 9e; Eq. 8):

$$\begin{aligned} \hat{\eta}_i(t) &= y_i(t) - \delta\overline{y}(t) \\ \hat{\eta}_i(t) &\xrightarrow{\mathcal{F}} H_i(\omega,T) \text{ with } 0 \le t \le T \\ \hat{G}_{\eta\eta}(\omega,T) &= \frac{2}{T}\sum_{i=1}^{N} w_i |H_i(\omega,T)|^2 \end{aligned} \tag{8}$$

Here, δ (in this and subsequent equations) can be set to 0 if the power of the target response is much smaller than the background noise, or to 1 if the power of the target response is expected to be comparable to the background noise. Similarly, the DOF of the variance of the average response (v_y) within a time window of interest $T' \le T$ are given by:

$$v_y = v\frac{T'}{T} \tag{9}$$

[Figure 9 about here.]

The variance of the BN ($\hat{\sigma}^2_{\mathrm{BN}_i}$), used to compute the weights, can be estimated either as the average variance of P fixed points across a number of epochs (N_k), i.e., the across-epochs approach (Eq. 10; employed in the Fmpi-WAE approach), or from the variance within each epoch, i.e., the within-epochs approach (Eq. 11; employed in the Fmpi-WWE approach).

$$\begin{aligned} \overline{y}_{N_k}(t) &= \frac{1}{N_k}\sum_{i=k_0}^{k_0+N_k-1} y_i(t) \\ \hat{\eta}_i(n\Delta t) &= y_i(n\Delta t) - \overline{y}_{N_k}(n\Delta t) \\ \hat{\sigma}^2_{\mathrm{BN}_k} = \frac{1}{P}\sum_{n=1}^{P}\frac{1}{N_k-1}\sum_{i=k_0}^{k_0+N_k-1}\hat{\eta}_i(n\Delta t)^2 &= \frac{1}{P}\sum_{n=1}^{P}\hat{\sigma}^2_{\mathrm{BN}_{kn}} \\ \hat{\sigma}^2_{\mathrm{BN}_i} &= \hat{\sigma}^2_{\mathrm{BN}_k} \text{ for } i \in [k_0,\ldots,k_0+N_k-1] \end{aligned} \tag{10}$$

$$\begin{aligned} \overline{y}(t) &= \frac{1}{N}\sum_{i=0}^{N-1} y_i(t) \\ \hat{\eta}_i(t) &= y_i(t) - \delta\overline{y}(t) \\ \hat{\sigma}^2_{\mathrm{BN}_i} &= \frac{1}{P-1}\sum_{n=1}^{P}\left(\hat{\eta}_i(n\Delta t) - \overline{\hat{\eta}_i(n\Delta t)}\right)^2 \end{aligned} \tag{11}$$

where k_0 (Eq. 10) is the starting epoch of a block of N_k epochs, and Δt is the time interval between the P tracked samples. In this study, P always corresponded to the number of samples available within an epoch, i.e., $\Delta t = 1/f_s$, with f_s being the sampling rate.

The choice of the noise variance estimation depends on the time scale of the response and the underlying noise dynamics. If the EP is measured using many (thousands of) epochs of short duration (< 20 ms) at a high repetition rate (as in auditory brainstem responses (ABRs)), where the dynamic fluctuations of the BN are slow relative to the repetition rate (e.g., slow myogenic artefacts lasting hundreds of milliseconds), the across-epochs approach (Eq. 10) is more accurate. This is because the total number of data points used to estimate the common noise variance within a block of epochs is greater than if only the data of a single epoch were used, thereby better capturing the dynamic fluctuations of the BN for weight estimation. In contrast, for cortical EPs, which typically consist of a few hundred epochs of long duration (> 800 ms) obtained at a low repetition rate, the variance of the BN can be accurately estimated within each epoch.

Frequency-domain EP

Several types of physiological responses (e.g., auditory steady-state response (ASSR), frequency following responses (FFRs), steady-state visual evoked potentials (SSVEPs), and otoacoustic emissions (OAEs)) evoke frequency-specific activity. In such cases, it is necessary to determine whether the power of the frequency components of interest is significantly greater than the power of the underlying noise. The Fmpi can be conveniently applied for this purpose. Two approaches are provided: one which is suitable for narrow-to-wide band frequency analysis, and another suitable for sparse or single-frequency component analysis.

Fmpi in the frequency-domain—narrow-band

Consider a narrow-band, weighted-averaged signal (Eq. 3) with frequencies of interest $\omega_j \in [w_a, w_b, \ldots, w_m]$. The post-average variance (Supplementary information) is given by Eq. 12, whilst the variance of the BN within each epoch is given by Eq. 13.

$$\begin{aligned} \overline{y}(t) &\xrightarrow{\mathcal{F}} Y(\omega_i, T) \text{ with } 0 \le t \le T \\ \hat{\sigma}_y^2 &= \frac{1}{P(P-1)} \sum_{\omega_i > 0} 2|Y(\omega_i, T)|^2 \end{aligned} \tag{12}$$

$$\begin{aligned} \hat{\eta}_i(t) &= y_i(t) - \delta\overline{y}(t) \\ \hat{\eta}_i(t) &\xrightarrow{\mathcal{F}} H_i(\omega_j, T) \\ \hat{\sigma}_{\mathrm{BN}_i}^2 &= \frac{1}{P(P-1)} \sum_{\omega_j} 2|H_i(\omega_j, T)|^2 \end{aligned} \tag{13}$$

The frequency-specific residual noise is obtained via Eq. 4 and the F-value via Eq. 1.
The DOF of the F-value are obtained from the weighted-averaged PSD (Eq. 8) as shown below:

$$v_y = \frac{2\left[\sum_{\omega_j} \hat{G}_{\eta\eta}(\omega_j, T)\right]^2}{\sum_{\omega_j} \hat{G}_{\eta\eta}(\omega_j, T)^2} \tag{14}$$

$$v_{\mathrm{RN}} = \frac{2\left[\sum_{\omega_j} \hat{G}_{\eta\eta}(\omega_j, T)\right]^2}{\sum_{\omega_j} \hat{G}_{\eta\eta}(\omega_j, T)^2} \frac{\left[\sum_{i=1}^{N} w_i\right]^2}{\sum_{i=1}^{N} w_i^2} = \frac{2\left[\sum_{\omega_j} \hat{G}_{\eta\eta}(\omega_j, T)\right]^2}{\sum_{\omega_j} \hat{G}_{\eta\eta}(\omega_j, T)^2} N_e \tag{15}$$

Fmpi in the frequency-domain—sparse-frequency analysis

In cases where only a few M sparsely distributed frequencies of interest $\omega_j \in [w_a, w_b, \ldots, w_m]$, such as the fundamental frequency and several harmonics, are relevant, it is necessary to assess whether the pooled power of these frequencies is significantly greater than the power of the underlying background noise at the same frequencies. This can be achieved by assuming that the noise at the frequencies of interest is similar to that at their neighbouring frequencies. Under this assumption, the pooled variance of the BN can be more accurately estimated—with increased DOF—by averaging the power of N_{ω_j} frequencies in the neighbourhood of each frequency of interest (Eq. 16).

$$\begin{aligned} \hat{\eta}_i(t) &= y_i(t) - \delta\overline{y}(t) \\ \hat{\eta}_i(t) &\xrightarrow{\mathcal{F}} H_i(\omega_j, T) \\ \overline{|H_i(\omega_j, T)|^2} &= \frac{1}{N_{\omega_j}} \sum_{\omega_k = \omega_j - \Delta_{\omega_j}}^{\omega_j + \Delta_{\omega_j}} |H_i(\omega_k, T)|^2 \\ \sigma_{\mathrm{BN}_i}^2 &= \frac{1}{P(P-1)} \sum_{\omega_j} 2\overline{|H_i(\omega_j, T)|^2} \text{ with } \omega_j > 0 \end{aligned} \tag{16}$$

The frequency-specific residual noise variance is obtained using Eq. 4, the weighted-averaged signal variance using Eq. 12, the F-value using Eq. 1, the DOF of the weighted-averaged signal using Eq. 14, and the DOF of the post-average RN is given by Eq. 17.

$$v_{\mathrm{RN}} = \frac{2\left[\sum_{\omega_n} \hat{G}_{\eta\eta}(\omega_n, T)\right]^2}{\sum_{\omega_n} \hat{G}_{\eta\eta}(\omega_n, T)^2} \frac{\left[\sum_{i=1}^{N} w_i\right]^2}{\sum_{i=1}^{N} w_i^2} = \frac{2\left[\sum_{\omega_n} \hat{G}_{\eta\eta}(\omega_n, T)\right]^2}{\sum_{\omega_n} \hat{G}_{\eta\eta}(\omega_n, T)^2} N_e \tag{17}$$

with $\omega_n \in [\omega_1 - \Delta\omega_1, \ldots, \omega_1 + \Delta\omega_1, \omega_2 - \Delta\omega_2, \ldots, \omega_2 + \Delta\omega_2, \ldots, \omega_m - \Delta\omega_m, \ldots, \omega_m + \Delta\omega_m]$.

Futility

In many situations, it is desirable to estimate how much longer a measurement should continue—either to predict the total recording time required to detect an EP, or, alternatively, to terminate the recording earlier if there is insufficient evidence that an EP will be detected within a reasonable timeframe (i.e., futility stopping). The Fmpi can be used to predict and update the necessary number of epochs (N_s), based on the ongoing information obtained from N epochs during a recording. This estimation is given by Eq. 18 (see Sup. 1.9.4 for details).

$$\begin{aligned} \frac{\hat{\sigma}^2_{BN_{eq}}}{max\left(\hat{\sigma}^2_{RN}, \hat{\sigma}^2_y\right)} Fc_{v_y v'_{RN}} < N_s \leq NFc_{v_y v'_{RN}} \\ \text{with } \hat{\sigma}^2_{BN_{eq}} = N\hat{\sigma}^2_{RN} \\ v'_{RN} = \frac{v_{RN}}{N_e}(N_e + N_s - N) \end{aligned} \tag{18}$$

Here, N_s is the predicted number of epochs required to detect a response, based on the information obtained from the current N epochs—specifically, the current $\hat{\sigma}^2_{RN}$, the current DOF (v_y and v_{RN}), and the current effective number of epochs N_e. Fc is the critical F-value required to reject the null-hypothesis.
The right side on Eq. 18 reflects the upper limit imposed by the fact the variance of the averaged response $\hat{\sigma}^2_y$ cannot be smaller than the variance of the residual noise $\hat{\sigma}^2_{RN}$ (see Sup. 1.9.4 for details). This relationship can be exploited to estimate estimate the DOF of the residual noise (v'_{RN} in Eq. 18) at the predefined maximum number of epochs N_f—the futility threshold—and, consequently, to determine the critical F-value (Fc) required to reject the null hypothesis. Importantly, the critical F-value depends on both v_y and v'_{RN}.
Under the assumption that recording conditions remain constant, this framework enables prediction of whether the number of epochs required to obtain a statistically significant response will exceed the futility threshold. Measurements can therefore be terminated once the condition $N_s > N_f$ is met.
Notably, in the absence of a neural response, $N_s = NFc$; that is, the number of epochs required to detect a response will increase by a factor of NFc, where the critical F-value is determined by the spectral characteristics—colour—of the individual underlying noise.

Fmp statistics

The original Fmp method[24,25,29] employing weighted averaging across epochs (Fmp-WAE) utilised an F-value calculation that was identical to that of the Fmpi-WAE. The sole distinction was that, in the original Fmp-WAE, the degrees of freedom (DOF) were fixed at $v_y = 5$ and $v_{RN} = N$[24,25,29]. For ASSRs, the Fmp-WAE was computed in the same manner as the frequency-based Fmpi-WAE; however, the degrees of freedom were unweighted, specifically $v_y = 2N_\omega$ and $v_{RN} = N$, where N_ω denotes the number of frequency bins included in the analysis.

Hotelling-T2 and modified T2 statistics

The Hotelling-T2[68] is a multivariate method that has been widely employed for the detection of EPs in both the time and frequency domains[13,36,40,41,43,44]. The Hotelling-T2 statistic is defined by the following equations:

$$\begin{aligned} \hat{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^\top \\ T^2 = N(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^\top \hat{\Sigma}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \\ F = \frac{N-P}{P(N-1)}T^2 \end{aligned} \tag{19}$$

Here, F follows a Fisher–Snedecor distribution with DOF P and $N - P$, where N is the number of epochs and P the number of features contained in the mean feature vector $\overline{\boldsymbol{x}}$. The mean feature vector $\overline{\boldsymbol{x}}$ is compared with $\boldsymbol{\mu} = 0$; $\hat{\Sigma}$ is the sample covariance matrix, and $\top$ denotes the transpose operator. The P mean features in $\overline{\boldsymbol{x}}$ correspond to the averages of across the N columns of the feature matrix $\boldsymbol{x}_{P\times N}$.
In this study, the feature matrix was constructed using attributes computed either in the time domain (i.e., several amplitude means within a predefined analysis time window) or in the frequency domain (i.e., real and imaginary Fourier coefficients from a Fourier transform applied to the analysis time window)[36,40,41,43,44]. To distinguish between these approaches, the acronym HT2 was used for time-domain-based features, and HT2-f for frequency-domain-based features. For frequency-domain-based features, either the full sample covariance matrix was used (HT2-f), or a version in which all off-diagonal elements were set to zero (HT2-f-diag). For the latter, the null distribution was estimated either under the

assumption that the DOF are the same as with the full sample covariance matrix, or empirically by bootstrapping the T^2 statistic from Eq. 19 (see Bootstrapping)—the modified T2 statistics test (T2-f-diag[44]).
It should be noted that 1) the F-value in Eq. 19 depends on both the number of epochs (N) and the number of features (P), 2) that the statistical power of the Hotelling-T2 decreases as the number of features increases—the so-called 'small n large p' problem[69]—, and 3) that the method becomes invalid when $P >= N$.

Q-sample

In addition to Fisher–Snedecor based detection methods, we employed the modified Q-sample test[33,34,37,38,43,44]. The Q-sample test is a non-parametric approach that calculates the W test statistic from q frequency bins obtained for each of the N epochs, as shown below:

$$
\begin{aligned}
W = \frac{2}{N} * \frac{2^2}{q^2(q+1)^2} \sum_{k=1}^{q} \left(C_k^2 + S_k^2\right) \\
C_k = \sum_{i=1}^{N} \alpha_{ik} cos(\beta_{ik}) \\
S_k = \sum_{i=1}^{N} \alpha_{ik} sin(\beta_{ik})
\end{aligned}
\tag{20}
$$

Here, α_{ik} may represent either the original spectral amplitude of the frequency bin or a ranked version thereof, and β_{ik} may be either the original spectral phase or a ranked version. In this study, the original phase was retained, and the amplitude of the frequency bins was ranked uniformly within the range $\alpha_{ik} \in 1, 2, 3, ..., Nq$[43,44]. The ranking strategy employed in the Q-Sample test aims to avoid excessive weights to epochs with large artefacts, especially when the number of epochs is reduced. The null distribution was obtained empirically via bootstrapping (see Bootstrapping).

Simulations

To evaluate the performance of the proposed Fmpi method, we conducted realistic simulations of EPs. These simulations utilised a template response (Figure 1a, target) embedded in $1/f$ pink noise, which closely approximates human brain background activity[60], under both stationary and non-stationary noise conditions (the latter including eye-blinking artefacts). Pink noise was generated by filtering a Gaussian white noise signal using a finite impulse response (FIR) filter with a $1/f^2$ transfer function. Eye-blink artefacts were simulated by randomly assigning unitary impulses over time at an average interval of 2 s (SD = 1.4 s). These impulses were subsequently convolved with a human eye-blink template[70] to generate the artefact waveform.
For cortical simulations, the power of the stationary background noise was adjusted such that the final signal-to-noise ratio (SNR) within the analysis filter bandwidth (2-30 Hz) was 3 dB after 200 epochs (Figure 1a, Signal + Noise). For ASSR simulations, the same template signal was convolved with a train of unitary impulses presented at a rate of 41 Hz, producing an ASSR with energy at 41 Hz and its harmonics (Sup. Figure 1). Each simulation comprised 80 epochs of 1 second duration, resulting in a final SNR of ≈ 12 dB under stationary noise. Simulations were carried out twice: once with the neural response included, and once without, using identical noise in both cases (Figure 1a, Noise).

Recordings

To test the Fmpi detector, we analysed data from multiple studies, encompassing ABRs, ASSRs, and cortical recordings. These datasets included both research and clinical recordings from adults with normal and impaired hearing, as well as from infants. Detailed descriptions of the datasets are provided in Table 1.

Processing

To compare the performance of all detectors, identical datasets were processed using each detection method. These included the Fmpi with standard average (Fmpi-SA; unweighted average), Fmpi with within-epochs (Fmpi-WWE) or across-epochs (Fmpi-WAE) weighted averaging, Fmp-WAE[29,33–38,40,41,43,44], Hotelling-T2 test using features from both the time (HT2) and frequency (HT2-f and HT2-f-diag) domains, and the fully bootstrapped T2-f-diag[44] and Q-Sample methods.
For ABR analyses, all data were downsampled to 8000 Hz, with epochs of 20 ms duration and an analysis window of 15 ms (either 2-17 ms or 5-20 ms, depending on stimulus characteristics). Cortical recordings were downsampled to 500 Hz, with epochs of ≈ 1 s and an analysis window of 500 ms (50 to 550 ms).
ABR data were band-pass filtered between 68 and 1350 Hz, and cortical responses between ≈ 2 and 30 Hz (Kaiser linear phase filter, 65 dB stop-band attenuation, 1 Hz transition between pass and stop bands). The high-pass cutoff frequency

was set to the inverse of the analysis window length to reduce bias in F-value estimation for the Fmp-WAE and Fmpi methods[24,29,71].

To capture the temporal dynamics of each method as a function of available data, recordings were analysed in incremental blocks of epochs: 200 for ABR, 20 for cortical, and 5 for ASSR recordings. To maximise the number of independent measurements per block, all possible independent blocks within each recording were included. For example, an ABR recording with 4000 epochs yielded 20 independent blocks of 200 epochs, 10 blocks of 400 epochs, 6 blocks of 600 epochs, 5 blocks of 800 epochs, and so forth. Due to the high computational cost of bootstrapping with ABR recordings, the number of included epochs was limited to 4000.

Block sizes for weight estimation with Fmp-WAE and Fmpi-WAE were 200, 20, 5 and epochs for ABR, cortical, and ASSR measurements, respectively. The δ parameter in equations Eq. 8, Eq. 11, Eq. 13, and Eq. 16 was set to zero, as the energy of the BN was typically more than 20 dB greater than that of the EPs analysed in this study.

To evaluate the accuracy and robustness of the detection methods under conditions of non-stationary noise, the rejection of noisy epochs was deliberately omitted.

Feature selection

Time-domain detectors utilised various features, all extracted within a consistent analysis window. For the HT2 method, amplitude means served as features, comprising either 40 (375 µs) amplitude means for ABRs[40], or 10 (50 ms) amplitude means for cortical analyses[36,41,43]. For cortical responses, frequency-domain detectors employed frequency bins between ≈ 2 and 8 Hz (yielding 4 bins or 8 features)[44], and between ≈ 67 and 600 Hz (9 bins or 18 features) for ABRs.

For ASSRs, all detectors incorporated the same features: the real and imaginary components of the frequencies of interest (i.e., the evoking frequency and its harmonics). In ASSR simulations and in all studies except one (study 10, Table 1) three harmonics (six features) were used—the fundamental and the next two harmonics. In study 10, twelve harmonics (24 features) were employed[34,37,38]. The choice between three or twelve harmonics was guided by two considerations. First, as indicated before, the performance of the Hotelling-T2 can be compromised as the number of features increases ('small n large p' problem)[69]. Second, the averaged ASSR waveforms from studies 8 and 9 (Figure 6a and Figure 7a) exhibited prominent peaks primarily at the first three harmonics. Thus, using three harmonics was intended to preserve good performance for HT2-f and HT2-f-diag while retaining relevant information. However, since previous studies employing the Q-Sample method included between six and twelve harmonics, twelve harmonics were used in study 10—accepting a reduction in performance for frequency-based Hotelling-T2 methods—to maintain comparability with Q-Sample analyses[37,38].

The bandwidth for noise estimation in the frequency-domain Fmpi method was set to 4 Hz ($\Delta\omega_j$ = 2 Hz in Eq. 16).

No response data

Due to the limited availability of ABR and cortical recordings obtained without stimulation—which are essential for assessing the sensitivity of different detection methods—we generated no-response datasets from the original recordings. First, the evoked response was estimated for each recording (Figure 9d) using all available epochs. This estimated response was then subtracted from the raw data (Figure 9a), at each event onset. Subsequently, the original event times were replaced by random event times (within the time range of the data) to ensure that any residual evoked response was eliminated when averaging the epochs (i.e., incoherent averaging). This procedure also prevented the final average from converging to zero as the number of epochs approached the total used to generate the subtracted evoked response. Simulations and analysis of real data (not shown) confirmed that this approach preserved the properties of the underlying noise—such as power, spectral shape, and correlation—whilst effectively removing the evoked response. For ASSRs, the evoked response was not subtracted; instead, analyses were performed on neighbouring frequency bins where no response was expected.

Bootstrapping

As previously described, we utilised the non-parametric Q-sample[33,34,37,38,43] and the modified T2 statistic test[44], with null distributions estimated empirically via bootstrapping[59,54]. First, the evoked response was subtracted from the continuous data (as outlined above). Subsequently, new random event times—equal in number to the original events and uniformly distributed within the original event time window—were generated. These randomised event times were applied to the continuous data, and either the W or T2 statistics were calculated from the resulting set of surrogate epochs. For each tested point within each recording, a total of 1000 test values were bootstrapped to construct the null distribution[43].

Performance metrics

Detector performance was evaluated using several metrics, including detection bias, false response detection rate, response detection rate, receiver operating characteristic curve (ROC), area under the curve (AUC), and average area under the curve (AAUC). Detection bias was quantified from the p-value distribution obtained from the no-response datasets (Figure 1b), which should be uniformly distributed for an unbiased detector[61]. Observed sensitivity was calculated by integrating the p-value distribution (Figure 2a) and bias was defined as the mean absolute difference between nominal and observed sensitivity (Figure 1b), assessed either across the full sensitivity range (0-100%) or within the clinically relevant range

(95-100%). False response detection rates (from no-response data) and response detection rates (from original datasets) were determined as a function of time (in epochs) at a nominal sensitivity of 95% (Figure 1c&d), by pooling all available detections for each participant within each study.

ROCs were constructed for different testing times (in epochs; Figure 2b&c) using sensitivity and specificity rates obtained for all possible nominal sensitivities or specifically for the clinically relevant range (95-100%). The AUC as a function of time (Figure 2d) was calculated as the area under each ROC and normalised by the total possible area, i.e., by 1 for all possible nominal sensitivities—with a corresponding chance level of 0.5—or by 0.05 for the clinically relevant range of nominal sensitivities (95-100%)—with a corresponding chance level of 0.025. Finally, the AAUC (Figure 2e), representing the overall performance of each detector, was estimated by averaging the AUC across all epochs.

All EEG simulations and recordings were processed offline using pEEGy v2.0.0, an open-source Python 3–based analysis package[72].

Statistical analysis

All statistical analyses were conducted using linear mixed-effects (LME) models[73]. In all models, the dependent variable was the AAUC, the independent factor was the *detector* type, and the random factor was the *subject*. For the simulations, the 1000 runs were grouped into blocks of 20, with these groups treated as the random factor in the models. The DOF were estimated using the Satterthwaite approximation[74]. Conditional, random, and marginal residuals of the LME models were inspected visually[75].
When outliers were identified (typically fewer than 2% based on the 1.5 inter-quartile range criterion), they were removed and the LME model was refitted. Conditional, random, and marginal residuals of the models were checked by visual inspection. Most outliers originated from the Fmp-WAE and HT2-f-diag detectors, which exhibited lower performance compared to other detectors. Post hoc comparisons were performed using Holm's family-wise error rate (FWER) correction[76].
All statistical analyses were performed using the R software package[77], with the significance level set at $\alpha = 0.05$.

Data availability

The raw electroencephalography datasets analysed in this study originate from multiple independent studies and are subject to ethical, institutional, and consent-related restrictions. As a result, the raw data are not publicly available. Access requests should be directed to the investigators responsible for each dataset, as listed in Table 1, and will be assessed in accordance with the relevant data-sharing agreements. Derived data supporting the findings are available from the corresponding author upon reasonable request.

Code availability

All methods described in this study are implemented in pEEGy, an open-source Python package developed for the analysis and simulation of electroencephalography data and evoked responses. The pEEGy source code is publicly available and maintained under an open-source licence. The version used for the analyses reported here is archived with a persistent digital object identifier (10.5281/zenodo.19512653), enabling reproducibility of the results. Documentation and usage examples are provided with the software.

Acknowledgments

This study was supported by the Interacoustics Research Unit. The authors thank Linsey Van Yper, Lisbeth Birkelund Simonsen, Mats Rekswinkel, Sinnet Greve Bjerge Kristensen, Guy Lightfoot, Macarena Bowen, and Anisa Visram for allowing us to analyse their data. We also thank James Harte and Bue Kristensen for their support and inspiring discussions about this study.

Author contributions

JU conceptualized and developed the novel Fmpi method. JU, MC, DS, and SL defined and conceptualised comparisons with other detection methods. JU analysed both simulations and neurophysiological data and wrote the manuscript with input from all authors.

Declaration of interests

JU is listed as the inventor of a patent (US20250057465A1). The patent is related to the Fmpi method described in this article. Co-author SL owns stocks in Demant A/S, which owns Interacoustics A/S. The authors declare that the research was conducted in the absence of any other commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. *World report on hearing* (2021). Geneva: World health organization.
2. Starr, A., Picton, T. W., Sininger, Y., Hood, L. J., and Berlin, C. I. (1996). "Auditory neuropathy". *Brain* 119, pp. 741–753.
3. Matthies, C. and Samii, M. (1997). "Management of Vestibular Schwannomas (Acoustic Neuromas): The Value of Neurophysiology for Intraoperative Monitoring of Auditory Function in 200 Cases". *Neurosurgery* 40, p. 459.
4. Mehta, K., Watkin, P., Baldwin, M., Marriage, J., Mahon, M., and Vickers, D. (2017). "Role of Cortical Auditory Evoked Potentials in Reducing the Age at Hearing Aid Fitting in Children With Hearing Loss Identified by Newborn Hearing Screening". *Trends in Hearing* 21, pp. 1–16.
5. Lascano, A. M., Lalive, P. H., Hardmeier, M., Fuhr, P., and Seeck, M. (2017). "Clinical evoked potentials in neurology: a review of techniques and indications". *J Neurol Neurosurg Psychiatry* 88, pp. 688–696.
6. Jewett, D. L. and Williston, J. S. (1971). "Auditory-evoked far fields averaged from the scalp of humans." *Brain : a journal of neurology* 94, pp. 681–96.
7. Martin, B. A. and Boothroyd, A. (2000). "Cortical, auditory, evoked potentials in response to changes of spectrum and amplitude". *J. Acoust. Soc. Am.* 107, pp. 2155–2161.
8. Cigánek, L. (1961). "The EEG response (evoked potential) to light stimulus in man". *Electroencephalography and Clinical Neurophysiology* 13, pp. 165–172.
9. Taylor, M. J. and McCulloch, D. L. (1992). "Visual Evoked Potentials in Infants and Children". *Journal of Clinical Neurophysiology* 9, p. 357.
10. Aminoff, M. J. and Goodin, D. S. (1994). "Visual Evoked Potentials". *Journal of Clinical Neurophysiology* 11, p. 493.
11. Colebatch, J. G. and Halmagyi, G. M. (1992). "Vestibular evoked potentials in human neck muscles before and after unilateral vestibular deafferentation". *Neurology* 42, pp. 1635–1635.
12. Galambos, R., Makeig, S., and Talmachoff, P. J. (1981). "A 40-Hz auditory potential recorded from the human scalp." *Proc Natl Acad Sci U S A* 78, pp. 2643–2647.
13. Picton, T. W., John, M. S., Dimitrijevic, A., and Purcell, D. W. (2003). "Human auditory steady-state responses." *Int. J. Audiol.* 42, pp. 177–219.
14. Undurraga, J. A., Haywood, N. R., Marquardt, T., and McAlpine, D. (2016). "Neural Representation of Interaural Time Differences in Humans—an Objective Measure that Matches Behavioural Performance". *JARO* 17, pp. 1–17.
15. Undurraga, J. A., Luke, R., Yper, L. V., Monaghan, J. J. M., and McAlpine, D. (2024). "The neural representation of an auditory spatial cue in the primate cortex". *Current Biology* 34, 2162–2174.e5.
16. Undurraga, J. A., Yper, L. V., Bance, M., McAlpine, D., and Vickers, D. (2020a). "Neural Encoding of Spectro-Temporal Cues at Slow and Near Speech-Rate in Cochlear Implant Users". *Hearing Research*, p. 108160.
17. Regan, D. (1977). "Steady-state evoked potentials". *J. Opt. Soc. Am., JOSA* 67, pp. 1475–1489.
18. Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottereau, B. R., and Rossion, B. (2015). "The steady-state visual evoked potential in vision research: A review". *J Vis* 15, p. 4.
19. Snyder, A. Z. (1992). "Steady-state vibration evoked potentials: description of technique and characterization of responses". *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* 84, pp. 257–268.
20. Bell, S. L., Fox, L., and Bihi, R. I. (2010). "Vestibular evoked myogenic responses to amplitude modulated sounds". *The Journal of the Acoustical Society of America* 128, pp. 559–562.
21. Wastell, D. G. (1977). "Statistical detection of individual evoked responses: An evaluation of Woody's adaptive filter". *Electroencephalography and Clinical Neurophysiology* 42, pp. 835–839.
22. Mason, S. M., Su, A. P., and Hayes, R. A. (1977). "Simple online detector of auditory evoked cortical potentials". *Med. Biol. Eng. Comput.* 15, pp. 641–647.
23. Wicke, J. D., Goff, W. R., Wallace, J. D., and Allison, T. (1978). "On-line statistical detection of average evoked potentials: Application to evoked response audiometry (ERA)". *Electroencephalography and Clinical Neurophysiology* 44, pp. 328–343.
24. Elberling, C. and Don, M. (1984). "Quality estimation of averaged auditory brainstem responses." *Scandinavian audiology* 13, pp. 187–97.
25. Don, M., Elberling, C., and Waring, M. (1984). "Objective detection of averaged auditory brainstem responses." *Scandinavian audiology* 13, pp. 219–28.
26. Arnold, S. A. (1985). "Objective versus Visual Detection of the Auditory Brain Stem Response". *Ear and Hearing* 6, p. 144.
27. Victor, J. D. and Mast, J. (1991). "A new statistic for steady-state evoked potentials." *Electroencephalography and clinical neurophysiology* 78, pp. 378–88.
28. Champlin, C. A. (1992). "Methods for detecting auditory steady-state potentials recorded from humans". *Hearing Research* 58, pp. 63–69.
29. Don, M. and Elberling, C. (1994). "Evaluating residual background noise in human auditory brain-stem responses." *JASA* 96, pp. 2746–57.
30. Henning, G. and Husar, P. (1995). "Statistical detection of visually evoked potentials". *IEEE Engineering in Medicine and Biology Magazine* 14, pp. 386–390.
31. Valdes, J. L., Perez-Abalo, M. C., Martin, V., Savio, G., Sierra, C., Rodriguez, E., and Lins, O. (1997). "Comparison of Statistical Indicators for the Automatic Detection of 80 Hz Auditory Steady State Responses". *Ear and Hearing* 18, p. 420.
32. Liavas, A., Moustakides, G., Henning, G., Psarakis, E., and Husar, P. (1998). "A periodogram-based method for the detection of steady-state visually evoked potentials". *IEEE Transactions on Biomedical Engineering* 45, pp. 242–248.

33. Stürzebecher, E., Cebulla, M., and Wernecke, K. (1999). “Objective response detection in the frequency domain: comparison of several q-sample tests.” *Audiology & neuro-otology* 4, pp. 2–11.

34. Cebulla, M., Stürzebecher, E., and Elberling, C. (2006). “Objective Detection of Auditory Steady-State Responses: Comparison of One-Sample and q-Sample Tests”. *J Am Acad Audiol* 17, pp. 093–103.

35. Silva, I. (2009). “Estimation of Postaverage SNR from Evoked Responses Under Nonstationary Noise”. *IEEE transactions on biomedical engineering* 56, pp. 2123–2123.

36. Golding, M., Dillon, H., Seymour, J., and Carter, L. (2009). “The detection of adult cortical auditory evoked potentials (CAEPs) using an automated statistic and visual detection”. *International Journal of Audiology* 48, pp. 833–842.

37. Stürzebecher, E. and Cebulla, M. (2013). “Automated auditory response detection: Improvement of the statistical test strategy”. *International Journal of Audiology* 52, pp. 861–864.

38. Cebulla, M. and Stürzebecher, E. (2015). “Automated auditory response detection: Further improvement of the statistical test strategy by using progressive test steps of iteration”. *International Journal of Audiology* 54, pp. 568–572.

39. Patel, R., Janawadkar, M. P., Sengottuvel, S., Gireesan, K., and Radhakrishnan, T. S. (2017). “Effective Extraction of Visual Event-Related Pattern by Combining Template Matching With Ensemble Empirical Mode Decomposition”. *IEEE Sensors Journal* 17, pp. 2146–2153.

40. Chesnaye, M. A., Bell, S. L., Harte, J. M., and Simpson, D. M. (2018). “Objective measures for detecting the auditory brainstem response: comparisons of specificity, sensitivity and detection time”. *International Journal of Audiology* 57, pp. 468–478.

41. Bardy, F., Van Dun, B., Seeto, M., and Dillon, H. (2020). “Automated cortical auditory response detection strategy”. *International Journal of Audiology* 59, pp. 835–842.

42. Derzsi, Z. (2021). “Optimal Approach for Signal Detection in Steady-State Visual Evoked Potentials in Humans Using Single-Channel EEG and Stereoscopic Stimuli”. *Front. Neurosci.* 15.

43. Chesnaye, M. A., Bell, S. L., Harte, J. M., Simonsen, L. B., Visram, A. S., Stone, M. A., Munro, K. J., and Simpson, D. M. (2021a). “Efficient Detection of Cortical Auditory Evoked Potentials in Adults Using Bootstrapped Methods”. *Ear and Hearing* 42, pp. 574–583.

44. — (2023). “Modified T2 Statistics for Improved Detection of Aided Cortical Auditory Evoked Potentials in Hearing-Impaired Infants”. *Trends Hear* 27, p. 23312165231154035.

45. Alpsan, D. (1991). “Classification Of Auditory Brainstem Responses By Human Experts And Backipropagation Neural Networks”. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society Volume 13: 1991*. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society Volume 13: 1991, pp. 1425–1426.

46. Acır, N., Özdamar, Ö., and Güzeliş, C. (2006). “Automatic classification of auditory brainstem responses using SVM-based feature selection algorithm for threshold detection”. *Engineering Applications of Artificial Intelligence* 19, pp. 209–218.

47. Davey, R., McCullagh, P., Lightbody, G., and McAllister, G. (2007). “Auditory brainstem response classification: A hybrid model using time and frequency features”. *Artificial Intelligence in Medicine* 40, pp. 1–14.

48. Parvar, H., Sculthorpe-Petley, L., Satel, J., Boshra, R., D’Arcy, R. C. N., and Trappenberg, T. P. (2015). “Detection of event-related potentials in individual subjects using support vector machines”. *Brain Inf.* 2, pp. 1–12.

49. McKearney, R. M. and MacKinnon, R. C. (2019). “Objective auditory brainstem response classification using machine learning”. *International Journal of Audiology* 58, pp. 224–230.

50. Zhu, B., Coppola, G., and Shoaran, M. (2019). “Migraine classification using somatosensory evoked potentials”. *Cephalalgia* 39, pp. 1143–1155.

51. Yperman, J., Becker, T., Valkenborg, D., Popescu, V., Hellings, N., Wijmeersch, B. V., and Peeters, L. M. (2020). “Machine learning analysis of motor evoked potential time series to predict disability progression in multiple sclerosis”. *BMC Neurol* 20, p. 105.

52. Xing, J., Qiu, S., Ma, X., Wu, C., Li, J., Wang, S., and He, H. (2020). “A CNN-based comparing network for the detection of steady-state visual evoked potential responses”. *Neurocomputing* 403, pp. 452–461.

53. McKearney, R. M., Bell, S. L., Chesnaye, M. A., and Simpson, D. M. (2022). “Auditory Brainstem Response Detection Using Machine Learning: A Comparison With Statistical Detection Methods”. *Ear and Hearing* 43, pp. 949–960.

54. Lv, J., Simpson, D. M., and Bell, S. L. (2007). “Objective detection of evoked potentials using a bootstrap technique”. *Medical Engineering & Physics* 29, pp. 191–198.

55. Maris, E. and Oostenveld, R. (2007). “Nonparametric statistical testing of EEG- and MEG-data”. *Journal of Neuroscience Methods* 164, pp. 177–190.

56. Chesnaye, M. A., Bell, S. L., Harte, J. M., and Simpson, D. M. (2021b). “Controlling test specificity for auditory evoked response detection using a frequency domain bootstrap”. *Journal of Neuroscience Methods* 363, pp. 1–12.

57. Lightfoot, G., Georga, C., Ferm, I., and Beeby, R. (2023). “The Clinical Evaluation of a Modified Fmp in Infant Auditory Brainstem Response Tests”. *Ear & Hearing* 44, pp. 924–929.

58. Haile, L. M., Kamenov, K., Briant, P. S., et al. (2021). “Hearing loss prevalence and years lived with disability, 1990–2019: findings from the Global Burden of Disease Study 2019”. *The Lancet* 397, pp. 996–1009.

59. Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Monographs on statistics and applied probability. New York: Chapman & Hall.

60. He, B. J., Zempel, J. M., Snyder, A. Z., and Raichle, M. E. (2010). “The Temporal Structures and Functional Significance of Scale-free Brain Activity”. *Neuron* 66, pp. 353–369.

61. Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2008). “P-Values are Random Variables”. *The American Statistician* 62, pp. 242–245.

62. Cohen, J. (1960). “A Coefficient of Agreement for Nominal Scales”. *Educational and Psychological Measurement* 20, pp. 37–46.

63. Landis, J. R. and Koch, G. G. (1977). “The Measurement of Observer Agreement for Categorical Data”. *Biometrics* 33, pp. 159–174.

64. Sohier, H., Bardy, F., and Ching, T. Y. C. (2021). “Relationship between objective measures of hearing discrimination elicited by non-linguistic stimuli and speech perception in adults”. *Sci Rep* 11, p. 19554.

65. Ching, T. Y. C., Zhang, V. W., Ibrahim, R., Bardy, F., Rance, G., Van Dun, B., Sharma, M., Chisari, D., and Dillon, H. (2023). “Acoustic change complex for assessing speech discrimination in normal-hearing and hearing-impaired infants”. *Clinical Neurophysiology* 149, pp. 121–132.

66. Albers, C. (2019). “The problem with unadjusted multiple and sequential statistical testing”. *Nat Commun* 10, p. 1921.

67. De Cheveigné, A. and Simon, J. Z. (2008). “Denoising based on spatial filtering.” *J. Neurosci. Methods* 171, pp. 331–9.

68. Hotelling, H. (1931). “The generalization of "Student’s" ratio”. *Annals of Mathematical Statistics* 2, pp. 360–378.

69. Bai, Z. and Saranadasa, H. (1996). “Effect of High Dimension: By an Example of a Two Sample Problem”. *Statistica Sinica* 6, pp. 311–329.

70. Valderrama, J. T., Torre, A. d. l., and Dun, B. V. (2018). “An automatic algorithm for blink-artifact suppression based on iterative template matching: application to single channel recording of cortical auditory evoked potentials”. *J. Neural Eng.* 15, pp. 1–16.

71. McKearney, R. M., Bell, S. L., Chesnaye, M. A., and Simpson, D. M. (2023). “Optimising weighted averaging for auditory brainstem response detection”. *Biomedical Signal Processing and Control* 83, p. 104676.

72. Undurraga, J. (2026). *pEEGy*.

73. Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). “Fitting Linear Mixed-Effects Models Using lme4”. *Journal of Statistical Software* 67, pp. 1–48.

74. Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). “lmerTest Package: Tests in Linear Mixed Effects Models”. *J. Stat. Softw.* 82, pp. 1–26.

75. Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., and Makowski, D. (2021). “performance: An R Package for Assessment, Comparison and Testing of Statistical Models”. *JOSS* 6, p. 3139.

76. Russell V. Lenth (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. Version R package version 1.10.6-090001.

77. R Development Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: Foundation for Statistical Computing.

78. Elberling, C. and Wahlgreen, O. (1985). “Estimation of auditory brainstem response, ABR, by means of Bayesian inference.” *Scandinavian audiology* 14, pp. 89–96.

79. Satterthwaite, F. E. (1946). “An Approximate Distribution of Estimates of Variance Components”. *Biometrics Bulletin* 2, pp. 110–114.

80. Satterthwaite, F. E. (1941). “Synthesis of variance”. *Psychometrika* 6, pp. 309–316.

81. Welch, B. L. (1956). “On Linear Combinations of Several Variances”. *Journal of the American Statistical Association* 51, pp. 132–148.

82. — (1947). “The Generalization of ‘Student’s’ Problem when Several Different Population Variances are Involved”. *Biometrika* 34, pp. 28–35.

83. Donelan, M. and Pierson, W. J. (1983). “The sampling variability of estimates of spectra of wind-generated gravity waves”. *Journal of Geophysical Research: Oceans* 88 (C7), pp. 4381–4392.

84. Medina, J. R., Aguilar, J., and Diez, J. J. (1985). “Distortions Associated with Random Sea Simulators”. *Journal of Waterway, Port, Coastal, and Ocean Engineering* 111, pp. 603–628.

85. Young, I. R. (1986). “Probability Distribution of Spectral Integrals”. *Journal of Waterway, Port, Coastal, and Ocean Engineering* 112, pp. 338–341.

86. Elgar, S. (1987). “Bias of Effective Degrees of Freedom of a Spectrum”. *Journal of Waterway, Port, Coastal, and Ocean Engineering* 113, pp. 77–82.

87. Rodríguez, G., Guedes Soares, C., and Machado, U. (1999). “Uncertainty of the sea state parameters resulting from the methods of spectral estimation”. *Ocean Engineering* 26, pp. 991–1002.

88. Bendat, J. S. and Piersol, A. G. (2010). *Random data: analysis and measurement procedures*. 4th ed. Wiley series in probability and statistics. Hoboken, N.J: Wiley. 604 pp.

89. Van Yper, L. N., Undurraga, Jaime A., Faundez, J. P., and McAlpine, D. (2023). “Fast or Slow Rates? A Comparison of Two Electrophysiological Measures of Binaural Processing”. In: *Association for Research in Otolaryngology 46TH Annual MidWinter Meeting*. Association for Research in Otolaryngology 46TH Annual MidWinter Meeting. Orlando, FL.

90. Simonsen, L. B., Undurraga, J. A., Kressner, A. A., Dau, T., and Laugesen, S. (2025). “Auditory Change Complex Responses to Spectrotemporally Modulated Stimuli”. *Ear and Hearing* 46, pp. 1538–1549.

91. Rekswinkel, M., Undurraga, J. A., and Kressner, A. A. (2023). “Cortical and behavioural discrimination in response to vowel-like spectral cues”. In: *International Evoked Response Audiometry Study Group Biennial Symposium 2023 (IERASG23)*. Cologne, Germany.

92. Visram, A. S., Stone, M. A., Purdy, S. C., Bell, S. L., Brooks, J., Bruce, I. A., Chesnaye, M. A., Dillon, H., Harte, J. M., Hudson, C. L., Laugesen, S., Morgan, R. E., O’Driscoll, M., Roberts, S. A., Roughley, A. J., Simpson, D., and Munro, K. J. (2023). “Aided Cortical Auditory Evoked Potentials in Infants With Frequency-Specific Synthetic Speech Stimuli: Sensitivity, Repeatability, and Feasibility”. *Ear and Hearing*.

93. Stone, M. A., Visram, A., Harte, J. M., and Munro, K. J. (2019). “A Set of Time-and-Frequency-Localized Short-Duration Speech-Like Stimuli for Assessing Hearing-Aid Performance via Cortical Auditory-Evoked Potentials”. *Trends Hear* 23, pp. 1–17.

94. Kristensen, S. G. B., Laugesen, S., and Dau, T. (2023). “Simultaneous recording of auditory brainstem Responses at clinical discharge level”. In: *International Symposium on Auditory and Audiological Research (ISAAR)*. International Symposium on Auditory and Audiological Research (ISAAR). Denmark.

95. Elberling, C. and Don, M. (2010). “A direct approach for the design of chirp stimuli used for the recording of auditory brainstem responses.” *The Journal of the Acoustical Society of America* 128, pp. 2955–64.

96. Lightfoot, G. and Stevens, J. (2013). “Effects of Artefact Rejection and Bayesian Weighted Averaging on the Efficiency of Recording the Newborn ABR.” *Ear and hearing*, pp. 1–8.

97. Bowen, M., Undurraga, J. A., and McAlpine, D. (2018). “Are objective measures of temporal fine structure and auditory nerve correlated with listening performance in noise? Evidence for Hidden Hearing Loss”. In: *Association for Research in Otolaryngology 41TH Annual MidWinter Meeting*. Association for Research in Otolaryngology 41TH Annual MidWinter Meeting. San Diego, USA.

98. Undurraga, J. A., McAlpine, D., and Lienart, J. (2020b). “On the Relation Between Binaural Temporal Processing and Speech Understanding in Noise”. In: *Association for Research in Otolaryngology 43TH Annual MidWinter Meeting*. Association for Research in Otolaryngology 43TH Annual MidWinter Meeting. San Jose, USA.

99. Laugesen, S., Rieck, J. E., Elberling, C., Dau, T., and Harte, J. M. (2018). “On the Cost of Introducing Speech-Like Properties to a Stimulus for Auditory Steady-State Response Measurements”. *Trends in Hearing* 22, pp. 1–11.

SUPPLEMENTARY INFORMATION

For the next developments we will consider the following definitions of mean, μ, and variance, σ^2, for a random variable x:

$$\mu_x = E\{x\} = \int_{-\infty}^{\infty} x\rho(x)\,dx \tag{S1}$$

$$\sigma_x^2 = E\left\{(x-\mu_x)^2\right\} = \int_{-\infty}^{\infty} (x-\mu_x)^2 \rho(x)\,dx = \mu_x = E\left\{x^2\right\} - \mu_x^2 \tag{S2}$$

where $\rho(x)$ is the probability density function of the variable x.
Consider a χ^2 random variable with n DOF

$$\chi_n^2 = z_1^2 + z_2^2 + \ldots + z_n^2 \tag{S3}$$

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \tag{S4}$$

where x_i is random normally distributed variable and z_i is a standardised random variable with $\mu_{z_i} = 0$ and $\sigma_{z_i} = 1$.
The following properties for χ^2 random variables can be demonstrated.

$$\mu_{\chi^2} = E\left\{\chi_n^2\right\} = n \tag{S5a}$$

$$var(\chi_n^2) = \sigma_{\chi_n^2}^2 = E\left\{\left(\chi_n^2 - \mu_{\chi_n^2}\right)^2\right\} = 2n \tag{S5b}$$

$$\frac{(n-1)\hat{\sigma}_x^2}{\sigma_x^2} = \chi_{n-1}^2 \tag{S5c}$$

$$n = \frac{2\left[E\left\{\chi_n^2\right\}\right]^2}{E\left\{\left(\chi_n^2 - \mu_{\chi^2}\right)^2\right\}} \tag{S5d}$$

$$E\left\{a\chi_n^2 + b\chi_m^2\right\} = an + bm \tag{S5e}$$

$$var\left(a\chi_n^2 + b\chi_m^2\right) = a^2 var(\chi_n^2) + b^2 var(\chi_m^2) = 2(a^2 n + b^2 m) \tag{S5f}$$

Sup. 1.1. F-statistic

By definition, the F-statistic corresponds to the ratio between the variance of two independent random variables. Specifically for an EP, the F-statistic corresponds to the ratio between the variance of the averaged signal, across many epochs within a recording (σ_y^2)—which contains both the signal of interest (x) and residual noise (RN)—and variance of the estimated residual noise (σ_{RN}^2).
If the signal of interest and the background noise are independent processes, the F-ratio (this is derived later) can be expressed as

$$F = \frac{\hat{\sigma}_y^2}{\hat{\sigma}_{\mathrm{RN}}^2} = \frac{\hat{\sigma}_{(x+RN)}^2}{\hat{\sigma}_{\mathrm{RN}}^2} = \frac{\hat{\sigma}_x^2 + \hat{\sigma}_{\mathrm{RN}}^2}{\hat{\sigma}_{\mathrm{RN}}^2} = \widehat{SNR} + 1 \tag{S6}$$

Note that F is the ratio of two variances, thus it can be parametrised by a Fisher–Snedecor distribution as $F(x, v_y, v_{\mathrm{RN}})$ where v_y and v_{RN} are the DOF of the respective variances in Eq. S6.
The estimation of the residual noise variance $\hat{\sigma}_{\mathrm{RN}}^2$ depends on the type of average utilised.

Sup. 1.2. Response average

Let's consider the following time series

$$\begin{array}{ccccc} \mathbf{Y}(t) & & \mathbf{X}(t) & & \mathbf{BN}(t) \\ \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_N(t) \end{bmatrix} & = & \begin{bmatrix} X(t) \\ X(t) \\ \vdots \\ X(t) \end{bmatrix} & + & \begin{bmatrix} BN_1(t) \\ BN_2(t) \\ \vdots \\ BN_N(t) \end{bmatrix} \end{array} \tag{S7}$$

where $y_i(t)$ is the i-th recorded epoch of a total of N, $X(t)$ is the *true* response of interest, and $BN_i(t)$ is the underlying background noise present on each epoch. For simplicity, we will assume that both $X(t)$ and $BN_i(t)$ have zero mean.

The average response estimate can be expressed as

$$\begin{aligned}\overline{y}(t) &= \boldsymbol{w}^T\mathbf{Y}(t) = \boldsymbol{w}^T\mathbf{X}(t) + \boldsymbol{w}^T\mathbf{BN}(t) \\ &= X(t) + RN(t)\end{aligned} \tag{S8}$$

where $RN = \boldsymbol{w}^T\mathbf{BN}(t)$ corresponds to the post-averaged RN.
Using standard averaging, the weight vector $\boldsymbol{w}$ is given by

$$\boldsymbol{w}_s^T = \left[\frac{1}{N}\cdots\frac{1}{N}\right] \tag{S9}$$

whereas for weighted averaging based on the inverse of the variance[24,25,29], $\boldsymbol{w}$ is given by

$$\boldsymbol{w}_w^T = \frac{\left[\frac{1}{\sigma^2_{\mathrm{BN}_1}}\cdots\frac{1}{\sigma^2_{\mathrm{BN}_N}}\right]}{\left(\sum_{i=1}^{N}\frac{1}{\sigma^2_{\mathrm{BN}_i}}\right)} \tag{S10}$$

where $\sigma^2_{\mathrm{BN}_i}$ corresponds to the variance within each epoch.

Sup. 1.3. SNR and F-value

The SNR of the measurement can be written as

$$SNR = \frac{E\left\{\boldsymbol{w}^T\mathbf{X}(t)\mathbf{X}(t)^T\boldsymbol{w}\right\}}{E\left\{\boldsymbol{w}^T\mathbf{BN}(t)\mathbf{BN}(t)^T\boldsymbol{w}\right\}} = \frac{\sigma_x^2}{\sigma^2_{\mathrm{RN}}} \tag{S11}$$

As shown in Eq. S11, the SNR corresponds to the variance of the target neural signal (σ_x^2) and the variance of the RN (σ^2_{RN}).
The variance of the averaged recording (σ_y^2)—which includes both the target signal and the noise—can be written as

$$\begin{aligned}\sigma_y^2 &= E\{w^T\mathbf{Y}(t)\mathbf{Y}(t)^T w\} \\ &= E\left\{\left[\sum_i w_i Y_i(t)\right]^2\right\} \\ &= E\left\{\left[\sum_i w_i\left(X_i(t) + BN_i(t)\right)\right]^2\right\} \\ &= E\left\{\left[\sum_i w_i X_i(t) + w_i BN_i(t)\right]^2\right\} \\ &= E\left\{\sum_{i,j} w_i w_j X_i(t) X_j(t)\right\} + E\left\{\sum_{i,j} w_i w_j BN_i(t) BN_j(t)\right\} + \\ &\quad E\left\{\sum_{i,j} w_i w_j X_i(t) BN_j(t)\right\}\end{aligned} \tag{S12}$$

Under the assumption that the signal and the noise are independent processes ($E\{X_i(t)BN_j(t)\} = 0$), and that the noises at each epoch are also independent ($E\{BN_i(t)BN_j(t)\}_{i\neq j} = 0$) this can be simplified as

$$\sigma_y^2 = E\{w^T\mathbf{Y}(t)\mathbf{Y}(t)^T w\} = \sigma_x^2 + \sigma^2_{\mathrm{RN}} \tag{S13}$$

Therefore, Eq. S11 can be written as

$$SNR = \frac{\sigma_{\mathrm{x}}^2}{\sigma^2_{\mathrm{RN}}} = \frac{\sigma_y^2 - \sigma^2_{\mathrm{RN}}}{\sigma^2_{\mathrm{RN}}} = \frac{\sigma_y^2}{\sigma^2_{\mathrm{RN}}} - 1 = F - 1 \tag{S14}$$

Sup. 1.4. Estimation of the background noise variance across epochs

To compute the F-ratio we must find a reliable estimate of σ^2_{RN}, which in turn requires a reliable estimation of the underlying potentially non-stationary background noise variance $\sigma^2_{\mathrm{BN}_i}$. This can be achieved by means of two different methods.

The first method estimates $\sigma^2_{\mathrm{BN}_i}$ by tracking the variance of a single or multiple fixed time-points *across epochs*[24,78,29,35]. In this case, if the noise is a locally stationary ergodic process, the variance of such a local noise source can be calculated as

$$\hat{\sigma}^2_{\mathrm{BN}_k} = \frac{1}{P}\sum_{n=1}^{P}\frac{1}{N_k-1}\sum_{i=k_0}^{k_0+N_k-1}\left(y_i(n\Delta t)-\overline{y}(n\Delta t)\right)^2 = \frac{1}{P}\sum_{n=1}^{P}\hat{\sigma}^2_{\mathrm{BN}_{kn}} \tag{S15}$$

where P is the number of tracked points used to estimate the background noise, N_k is the number of epochs used to estimate the variance of each block, k_0 is the starting epoch of the block, and Δt is the time interval between the tracked samples.

This method assumes that the noise is locally stationary within each block of N_k epochs, that is, $\hat{\sigma}^2_{\mathrm{BN}_{k1}} = \hat{\sigma}^2_{\mathrm{BN}_{k2}} = \ldots = \hat{\sigma}^2_{\mathrm{BN}_{kP}}$.

Sup. 1.5. Estimation of the background noise variance within epoch

This second method estimates the underlying background noise variance $\sigma^2_{\mathrm{BN}_i}$ from the variance *within each epoch* as shown in Eq. S16.

$$\begin{aligned}\hat{\eta}_i(t) &= y_i(t)-\delta\overline{y}(t)\\ \hat{\sigma}^2_{\mathrm{BN}_i} &= \frac{1}{P-1}\sum_{n=1}^{P}\left(\hat{\eta}_i(n\Delta t)-\overline{\hat{\eta}_i(n\Delta t)}\right)^2\end{aligned} \tag{S16}$$

Here, δ is 0 or 1, pending on the assumption that the variance of the *true* target response $\sigma^2_x \ll \sigma^2_{BN_i}$, or not. In many real applications $\sigma^2_{BN_i}$ will 20 dB or more above σ^2_x. If that is the case, then the assumption will be true, and δ will be zero.

Sup. 1.6. Estimation of the residual noise variance

Regardless of the chosen method to estimate the background noise (Eq. S15 or Eq. S16), the variance of the residual noise σ^2_{RN} can be obtain as shown in Eq. S17.

$$\begin{aligned}\sigma^2_{\mathrm{RN}} = E\{w^T\mathbf{RN}(t)\,\mathbf{RN}(t)^T w\} &= E\left\{\left[\sum_i w_i BN_i(t)\right]^2\right\}\\ &= E\left\{\sum_i w_i^2 BN_i^2(t)+\sum_{i\neq j} w_i w_j BN_i(t)\,BN_j(t)\right\}\\ &= \sum_i E\left\{w_i^2 BN_i^2(t)\right\}+\sum_{i\neq j} E\left\{w_i w_j BN_i(t)\,BN_j(t)\right\}\\ &= \sum_{i=1}^{N} w_i^2\sigma^2_{BN_i}\end{aligned} \tag{S17}$$

Now, if weights are defined as in Eq. S10, Eq. S17 can be written as

$$\begin{aligned}\sigma^2_{\mathrm{RN}} = \sum_{i=1}^{N} w_i^2\sigma^2_{BN_i} &= \frac{1}{\left[\sum_{i=1}^{N}\frac{1}{\sigma^2_{\mathrm{BN}_i}}\right]^2}\cdot\sum_{i=1}^{N}\frac{1}{\sigma^2_{\mathrm{BN}_i}}\\ &= \left[\sum_{i=1}^{N}\frac{1}{\sigma^2_{\mathrm{BN}_i}}\right]^{-1}\end{aligned} \tag{S18}$$

Note that Eq. S17 can be understood as the summation of $\chi 2$ distributed random variables with different weights (Eq. S5e), that is

$$\sigma^2_{\mathrm{RN}} = \sum_{i=1}^{N} w_i^2\sigma^2_{BN_i} = \sum_{i=1}^{N} a_i\sigma^2_{BN_i} \tag{S19}$$

Sup. 1.7. Estimation of effective number of epochs

Averaging the data using any kind of weights will affect the way in which each epoch contributes to the data, so that the effective number of epochs will be lower than the total number of epochs (or epochs) N. However, given that Eq. S19 can be interpreted as the linear combination of several variances with unknown DOFs, we can estimate the effective sampling size (or DOF) using the Welch-Satterthwaite approximation[79–82,81].

By means of the properties of a χ^2 random variable (Eq. S5d), we can rewrite Eq. S19 and derive the DOF of a new distribution which shares the same first moment (μ_{χ^2}) and the same variance ($\sigma^2_{\chi^2}$), of σ^2_{RN}.

First, let's consider the case where $\sigma^2_{BN_i}$ is estimated by averaging the variance *across epochs* (Eq. S15). If all samples are independent and drawn from the same random process, we can simplify and write the following

$$\hat{\sigma}^2_{\mathrm{BN}_k} = \frac{1}{P}\sum_{n=1}^{P}\hat{\sigma}^2_{\mathrm{BN}_{kn}} \tag{S20}$$

Now, using Eq. S5c we can rewrite the variance of each of the n sample points as

$$\hat{\sigma}^2_{\mathrm{BN}_k} = \frac{1}{P}\sum_{n=1}^{P}\hat{\sigma}^2_{\mathrm{BN}_{kn}} = \frac{1}{P}\sum_{n=1}^{P}\frac{\chi^2_{(N_k-1)_n}\sigma^2_{\mathrm{BN}_{kn}}}{N_k-1} \tag{S21}$$

Keeping in mind that the noise is assumed to be a locally stationary ergodic processes (i.e., $\sigma^2_{\mathrm{BN}_{k1}} = \sigma^2_{\mathrm{BN}_{k2}} = \cdots = \sigma^2_{\mathrm{BN}_{kP}}$) and $\chi^2_{(N_k-1)_n}$ are n independent chi-squared variables with (N_k-1) DOF we obtain

$$\hat{\sigma}^2_{\mathrm{BN}_k} = \frac{\sigma^2_{\mathrm{BN}_k}}{P(N_k-1)}\sum_{n=1}^{P}\chi^2_{(N_k-1)_n} \tag{S22}$$

Next, considering the variance of the residual noise in the form of Eq. S19 combined with Eq. S22, we can rewrite $\hat{\sigma}^2_{\mathrm{RN}}$ as

$$\hat{\sigma}^2_{\mathrm{RN}} = \sum_{k=1}^{M} a_k\hat{\sigma}^2_{BN_k} = \sum_{k=1}^{M}\frac{a_k\sigma^2_{\mathrm{BN}_k}}{P(N_k-1)}\sum_{n=1}^{P}\chi^2_{(N_k-1)_n} \tag{S23}$$

Using Eq. S5f, Eq. S23, and Eq. S5b, we can estimate the variance of our estimate—$var\left(\hat{\sigma}^2_{\mathrm{RN}}\right)$—and represent it as the variance of a $\chi^2_{v_{\mathrm{RN}}}$ with unknown v_{RN} DOF

$$\begin{aligned} var\left(\hat{\sigma}^2_{\mathrm{RN}}\right) &= \\ var\left(\sum_{k=1}^{M}\frac{a_k\sigma^2_{\mathrm{BN}_k}}{P(N_k-1)}\sum_{n=1}^{P}\chi^2_{(N_k-1)_n}\right) &= \\ \sum_{k=1}^{M}\frac{a_k^2\sigma^4_{BN_k}\sum_{n=1}^{P}var\left(\chi^2_{(N_k-1)_n}\right)}{P^2(N_k-1)^2} &= \\ \sum_{k=1}^{M}\frac{a_k^2\sigma^4_{BN_k}2P(N_k-1)}{P^2(N_k-1)^2} &= \\ 2\sum_{k=1}^{M}\frac{a_k^2\sigma^4_{BN_k}}{P(N_k-1)} &= \\ E\left\{\left(\chi^2_{v_{\mathrm{RN}}}-\mu_{\chi^2_{v_{\mathrm{RN}}}}\right)^2\right\} &= 2v_{\mathrm{RN}} \end{aligned} \tag{S24}$$

Similarly, recalling Eq. S5a we can estimate the mean (or expected value) of our estimate—$E\left\{\hat{\sigma}^2_{\mathrm{RN}}\right\}$—and express it as the expected value of a $\chi^2_{v_{\mathrm{RN}}}$ with unknown v_{RN} DOF, i.e.,

$$\begin{aligned} E\left\{\hat{\sigma}^2_{\mathrm{RN}}\right\} &= E\left\{\sum_{k=1}^{M}a_k\frac{\sigma^2_{\mathrm{BN}_k}}{P(N_k-1)}\sum_{n=1}^{P}\chi^2_{(N_k-1)_n}\right\} \\ &= \sum_{k=1}^{M}a_k\frac{\sigma^2_{BN_k}\sum_{n=1}^{P}E\left\{\chi^2_{(N_k-1)_n}\right\}}{P(N_k-1)} \\ &= \sum_{k=1}^{M}a_k\sigma^2_{BN_k} = E\left\{\chi^2_{v_{\mathrm{RN}}}\right\} = v_{\mathrm{RN}} \end{aligned} \tag{S25}$$

Squaring Eq. S25 and dividing by Eq. S24, the unknown DOF can be estimated as

$$v_{\mathrm{RN}} = \frac{\left[\sum_{k=1}^{M} a_k \sigma_{BN_k}^2\right]^2}{\sum_{k=1}^{M} \frac{a_k^2 \sigma_{BN_k}^4}{P(N_k-1)}} \tag{S26}$$

Replacing the constants a_k by w_k^2, where w_k is given by Eq. S10, results in

$$\begin{aligned} v_{\mathrm{RN}} &= \frac{\left[\sum_{k=1}^{M} w_k^2 \sigma_{BN_k}^2\right]^2}{\sum_{k=1}^{M} \frac{w_k^4 \sigma_{BN_k}^4}{P(N_k-1)}} \\ &= P \frac{\left[\sum_{k=1}^{M} \frac{1}{\sigma_{BN_k}^2}\right]^2}{\sum_{k=1}^{M} \frac{1}{\sigma_{BN_k}^4 (N_k-1)}} \\ &= P \frac{\left[\sum_{k=1}^{M} w_k\right]^2}{\sum_{k=1}^{M} \frac{w_k^2}{N_k-1}} \end{aligned} \tag{S27}$$

where N_k is the number of epochs used to estimate each variance $\sigma_{BN_k}^2$, that is $\sum_{k=1}^{M} N_k = N$, with M being the number blocks.

Finally, if the weights are expanded to all N individual epochs, and the number of epochs within a block is constant $N_1 = N_2 = \cdots = N_k$, Eq. S27 can be approximated as

$$v_{\mathrm{RN}} = P \frac{\left[\sum_{k=1}^{M} w_k\right]^2}{\sum_{k=1}^{M} \frac{w_k^2}{N_k-1}} = P(N_k-1) \frac{\left[\sum_{k=1}^{M} w_k\right]^2}{\sum_{k=1}^{M} w_k^2} = \frac{P(N_k-1)}{N_k} \frac{\left[\sum_{i=1}^{N} w_i\right]^2}{\sum_{i=1}^{N} w_i^2} \approx P \frac{\left[\sum_{i=1}^{N} w_i\right]^2}{\sum_{i=1}^{N} w_i^2} = P N_e \tag{S28}$$

where N_e is the effective number of epochs estimated from the weights ($N_e \leq N$).

Following the same rationale, when $\sigma_{BN_i}^2$ is estimated *within each epoch* according to Eq. S16, the DOF of $\hat{\sigma}_{\mathrm{RN}}^2$ can be estimated as it follows

$$\hat{\sigma}_{\mathrm{RN}}^2 = \sum_{i=1}^{N} \frac{a_i \sigma_{BN_i}^2 \chi_{(P-1)_i}^2}{P-1} \tag{S29}$$

with variance

$$\begin{aligned} var\left(\hat{\sigma}_{\mathrm{RN}}^2\right) &= var\left(\sum_{i=1}^{N} \frac{a_i \sigma_{BN_i}^2 \chi_{(P-1)_i}^2}{P-1}\right) \\ &= \sum_{i=1}^{N} \frac{a_i^2 \sigma_{BN_i}^4 var\left(\chi_{(P-1)_i}^2\right)}{(P-1)^2} \\ &= \sum_{i=1}^{N} \frac{a_i^2 \sigma_{BN_i}^4 2(P-1)}{(P-1)^2} \\ &= 2\sum_{i=1}^{N} \frac{a_i^2 \sigma_{BN_i}^4}{P-1} = E\left\{\left(\chi_{v_{\mathrm{RN}}}^2 - \mu_{\chi^2}\right)^2\right\} = 2v_{\mathrm{RN}} \end{aligned} \tag{S30}$$

and expected value

$$
\begin{aligned}
E\left\{\hat{\sigma}^2_{\mathrm{RN}}\right\} &= E\left\{\sum_{i=1}^{N}\frac{a_i\sigma^2_{BN_i}\chi^2_{(P-1)_i}}{P-1}\right\} \\
&= \sum_{i=1}^{N} E\left\{\frac{a_i\sigma^2_{BN_i}\chi^2_{(P-1)_i}}{P-1}\right\} \\
&= \sum_{i=1}^{N} a_i\sigma^2_{BN_i} = E\left\{\chi^2_{v_{\mathrm{RN}}}\right\} = v_{\mathrm{RN}}
\end{aligned}
\tag{S31}
$$

Squaring Eq. S31 and dividing by Eq. S30, followed by substituting a_k by w_k^2 we obtain

$$
v_{\mathrm{RN}} = \frac{\left[\sum_{i=1}^{N} a_i\sigma^2_{BN_i}\right]^2}{\sum_{i=1}^{N}\frac{a_i^2\sigma^4_{BN_i}}{P-1}} = (P-1)\frac{\left[\sum_{i=1}^{N} w_i^2\sigma^2_{BN_i}\right]^2}{\sum_{i=1}^{N} w_i^4\sigma^4_{BN_i}}
\tag{S32}
$$

Finally, given that the weights correspond to Eq. S10, Eq. S32 can be rewritten as

$$
v_{\mathrm{RN}} = (P-1)\frac{\left[\sum_{i=1}^{N} w_i^2\sigma^2_{BN_i}\right]^2}{\sum_{i=1}^{N} w_i^4\sigma^4_{BN_i}} = (P-1)\frac{\left[\sum_{i=1}^{N} w_i\right]^2}{\sum_{i=1}^{N} w_i^2} \approx P\frac{\left[\sum_{i=1}^{N} w_i\right]^2}{\sum_{i=1}^{N} w_i^2} = PN_e
\tag{S33}
$$

Equation S28 and Eq. S33 are essentially identical and provide an estimate of the *effective sample size* when weighted averaging is used. This approximation assumes that the samples used to estimate the variances are independent, within and across epochs. Independency across epochs can be approximated by removing low frequency components (i.e., high-pass filtering the raw data) with a cut-off frequency higher than $1/T$. However, independency within epochs cannot be avoided either because of the properties of the background noise (coloured noise) and/or the use of non-ideal filters to reduce undesired noise. Therefore, the number of independent samples within epochs—P—used in Eq. S28 and Eq. S33, will be $\leq$ than the number of within-epoch samples used.

Sup. 1.8. Estimation of the degrees of freedom within epoch

Consider a continuous time series in which the underlying noise is an ideal white noise, i.e., all frequency components have the same energy. In this ideal case, the number of DOF will be only limited by the bandwidth of the noise. In the case of a discretised white noise signal, the number of DOF will limited by the number of discretised samples for a full bandwidth noise, and it will be less than this if the bandwidth of the noise is lower than the Nyquist frequency of the recording system. However, the ongoing brain activity is typically dominated by low-frequency energy, reflected by an uneven distribution of its spectral power—coloured noise—usually described as being $\propto 1/f$, with f being the frequency. A direct consequence of coloured noise is a higher level of correlation among samples which in turn reduces their independence as well as the effective DOF relative to a white noise.

This problem has been investigated in other fields and it forms the basis of the next steps to derive the effective DOF of the background neural activity present in each epoch[83–88].

Consider a random variable $y(t)$ with zero mean and standard deviation σ_y^2. The autocorrelation function of this signal is given by

$$
R_{yy}(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} y(t)\,y(t+\tau)\,dt = \sigma_y^2\delta(\tau)
\tag{S34}
$$

and its power spectral density is given by

$$
\begin{aligned}
S_{yy}(\omega) &= \int_{-\infty}^{\infty} R_{yy}(\tau)\,e^{-j\omega\tau}d\tau \\
R_{yy}(\tau) &= \int_{-\infty}^{\infty} S_{yy}(\omega)\,e^{j\omega\tau}d\omega
\end{aligned}
\tag{S35}
$$

Note that

$$
R_{yy}(\tau) = \int_{-\infty}^{\infty} S_{yy}(\omega)\,e^{j\omega\tau}d\omega = 2\int_{0}^{\infty} S_{yy}(\omega)\,e^{j\omega\tau}d\omega = \int_{0}^{\infty} G_{yy}(\omega)\,e^{j\omega\tau}d\omega,
\quad \text{with } G_{yy}(\omega) = 2S_{yy}(\omega)
\tag{S36}
$$

The power spectrum of an unlimited random process $y(t)$ is given by

$$G_{yy}(\omega) = 2\lim_{T\to\infty}\frac{1}{T}E\left\{|Y(\omega,T)|^2\right\} \tag{S37}$$

Therefore, an approximation for a time-limited signal is given by

$$\tilde{G}_{yy}(\omega,T) = \frac{2}{T}|Y(\omega,T)|^2 \tag{S38}$$

An important observation is that

$$\tilde{G}_{yy}(\omega,T) = \frac{2}{T}|Y(\omega,T)|^2 = \frac{2}{T}\left[Y_{re}^2(\omega,T) + Y_{im}^2(\omega,T)\right] \tag{S39}$$

where the real and imaginary part are both normally distributed with

$$\begin{aligned} E\{Y_{re}\} = E\{Y_{im}\} &= 0 \\ E\left\{Y_{re}^2\right\} = E\left\{Y_{im}^2\right\} &= \sigma_Y^2 \\ E\{Y_{re}Y_{im}\} &= 0 \end{aligned} \tag{S40}$$

By definition, Eq. S39 follows a Rayleigh distribution, i.e., a χ^2 distribution with two DOF, that is, $\tilde{G}_{yy}(\omega,T) \sim \chi_2^2$. Therefore, using Eq. S5c, we can write

$$\frac{\tilde{G}_{yy}(\omega,T)}{G_{yy}(\omega,T)} = \frac{\chi_n^2}{n}, \text{with } n = 2 \tag{S41}$$

The normalised error of $\tilde{G}_{yy}(\omega)$ is given by

$$\begin{aligned} e\left(\tilde{G}_{yy}(\omega,T)\right) &= \frac{\sigma\left(\tilde{G}_{yy}(\omega,T)\right)}{E\{\tilde{G}_{yy}(\omega,T)\}} = \frac{\sigma\left(\tilde{G}_{yy}(\omega,T)\right)}{E\{G_{yy}(\omega,T)\frac{\chi_n^2}{n}\}} \\ &= \frac{\sqrt{var\left(G_{yy}(\omega,T)\frac{\chi_n^2}{n}\right)}}{G_{yy}(\omega,T)} = \frac{\sqrt{2n}}{n} \end{aligned} \tag{S42}$$

where the expected value and variance terms were derived using Eq. S5a and Eq. S5b, respectively.
Now, using Eq. S42, we can derive the variance of the estimated power spectrum as

$$\sigma^2\left(\tilde{G}_{yy}(\omega,T)\right) = \frac{2}{n}\left[E\{\tilde{G}_{yy}(\omega,T)\}\right]^2 = \frac{2}{n}\left[G_{yy}(\omega,T)\right]^2 \tag{S43}$$

Thus, a better approximation of G_{yy} can be obtained by averaging several realisations of $\tilde{G}_{yy}$

$$\hat{G}_{yy}(\omega,T) = \frac{1}{N}\sum_{i=1}^{N}\tilde{G}_{yy_i}(\omega,T) = \frac{2}{NT}\sum_{i=1}^{N}|Y_i(\omega,T)|^2 \tag{S44}$$

where the normalised error is now given by

$$e\left(\hat{G}_{yy}(\omega,T)\right) = \frac{\sigma\left(\hat{G}_{yy}(\omega,T)\right)}{E\left\{\hat{G}_{yy}(\omega,T)\right\}} = \frac{\sigma\left(\hat{G}_{yy}(\omega,T)\right)}{G_{yy}(\omega,T)} = \frac{\sqrt{\frac{2nN}{N^2}}}{n} = \frac{1}{\sqrt{N}}, \text{ when n=2} \tag{S45}$$

Note that Eq. S44 can be also estimated using weighted average via Eq. S8.

$$\hat{G}_{yy}(\omega,T) = \sum_{i=1}^{N} w_i\tilde{G}_{yy_i}(\omega,T) = \frac{2}{T}\sum_{i=1}^{N} w_i|Y_i(\omega,T)|^2 = \frac{2}{T}\overline{|Y(\omega,T)|^2} \tag{S46}$$

Although Eq. S46 represent a continuous function of ω, the following derivations will assume that we are working with a discrete representation, i.e., ω_j, with $j \in [1, N_\omega]$.
To derive the DOF of a signal, we can now invoke Parseval's theorem which establishes that the power of a time-domain signal is equal to the sum of the power spectrum across frequencies, i.e., $Pw = \frac{1}{N_\omega}\sum_j G_{yy}(\omega_j,T)$, which for a zero-mean signal is equivalent to the variance of the signal.
Note that $G_{yy}(\omega_j,T)$ follows a χ^2 distribution, and so does its power $\sum_j G_{yy}(\omega_j,T)$.

Consequently, can invoke once more Eq. S5d to derive the DOF for the signal power as

$$v = \frac{2\left[E\left\{\chi_n^2\right\}\right]^2}{E\left\{(\chi_n^2 - \mu_{\chi^2})^2\right\}} = \frac{2\left[E\left\{\chi_n^2\right\}\right]^2}{var(\chi_n^2)} = \frac{2\left[E\left\{Pw\right\}\right]^2}{var(Pw)} = \frac{2\left[E\left\{\sum_{j=1}^{N_\omega} \hat{G}_{yy}(\omega_j, T)\right\}\right]^2}{var\left\{\sum_{j=1}^{N_\omega} \hat{G}_{yy}(\omega_j, T)\right\}} \tag{S47}$$

The expected mean of the power in Eq. S47 corresponds to

$$E\left\{\sum_{j=1}^{N_\omega} \hat{G}_{yy}(\omega_j, T)\right\} = \sum_{j=1}^{N_\omega} E\left\{\hat{G}_{yy}(\omega_j, T)\right\} = \sum_{j=1}^{N_\omega} E\left\{G_{yy}(\omega_j, T)\frac{\chi_2^2}{2}\right\} = \sum_{j=1}^{N_\omega} G_{yy}(\omega_j, T) \tag{S48}$$

The variance in the denominator of Eq. S47 can be estimated using Eq. S43

$$\begin{aligned} var\left\{\sum_{j=1}^{N_\omega} \hat{G}_{yy}(\omega_j, T)\right\} = var\left\{\sum_{j=1}^{N_\omega} G_{yy}(\omega_j, T)\frac{\chi_2^2}{2}\right\} &= \sum_{j=1}^{N_\omega} \left[G_{yy}(\omega_j, T)\right]^2 \frac{var\left\{\chi_2^2\right\}}{4} \\ &= \sum_{j=1}^{N_\omega} \left[G_{yy}(\omega_j, T)\right]^2 \end{aligned} \tag{S49}$$

Replacing Eq. S49 in Eq. S47 we obtain

$$v = \frac{2\left[\sum_{j=1}^{N_\omega} G_{yy}(\omega_j, T)\right]^2}{\sum_{j=1}^{N_\omega} \left[G_{yy}(\omega_j, T)\right]^2} \approx \frac{2\left[\sum_{j=1}^{N_\omega} \hat{G}_{yy}(\omega_j, T)\right]^2}{\sum_{j=1}^{N_\omega} \left[\hat{G}_{yy}(\omega_j, T)\right]^2} \tag{S50}$$

Eq. S50 thus provides an estimation for the DOF of the background activity within epochs of length T.
Note that the equivalent relationship in the continuum domain—useful to generate analytical solutions—is given by

$$v = \frac{2\left[\int_0^\infty G_{yy}(f)\,df\right]^2}{\frac{1}{T}\int_0^\infty \left[G_{yy}(f)\right]^2 df} \tag{S51}$$

with $f = \omega/2\pi$ and T being the time interval analysed.
This equivalence becomes apparent when approximating the integral using the rectangle rule, i.e., replacing $\int_0^\infty$ by $\sum_{j=1}^{N_f}$, df by Δf, and f by $j\Delta f$, with $\Delta f = 1/T$.
In Eq. S51 can be seen that whilst v is shaped by the PSD function, it scales linearly as a function of T.

Sup. 1.9. The Fmpi

As stated in Eq. S14, the statistical significance of the SNR is governed by the F-value, which in turn depends on the effective number of DOF v_y and v_{RN} of σ_s^2 and σ_{RN}^2, respectively. Different implementations based on the F-test[24,78,29,35] have used empirical estimations of the degrees of freedom of σ_s^2—based on the background activity recorded from humans—, whilst it has been assumed that the DOF of σ_{RN}^2 is given by the number of averaged epochs.
Whilst these approximations work for the average population, or an assumed worst-case scenario, they are unable to reflect the unique properties of the individual's underlying noise that determines true value of those parameters.
The Fmpi algorithm exploits the information extracted from the underlying noise to accurately estimate the effective number of DOF of it (Eq. S50), tailored—or individualised—to the EEG acquired in the individual, and uses this information to estimate the effective number of DOF of σ_{RN}^2. This is because Eq. S50 is an approximation of the effective number of independent samples within an epoch (P in Eq. S28 and Eq. S33).

Sup. 1.9.1. Fmpi for time-domain analysis

Consider that we have epochs of length T, and we are interested in testing whether a signal is present within a shorter analysis window of duration T', with $T' \leq T$.
To estimate the DOF of σ_y^2—the variance within the analysis window—v_y, we can make use of Eq. S50, which scales linearly as a function of the length of the time window chosen (Eq. S51).
Therefore, if σ_y^2 is computed from an analysis window T', the DOF (v_y; the numerator of the F ratio) can be estimated from the entire window of duration T (i.e., at a higher frequency resolution), and then scaled down by Eq. S52

$$v_y = v\frac{T'}{T} \tag{S52}$$

That is, v is estimated for the entire epoch of duration T using Eq. S50, with $\hat{G}_{yy}(\omega,T)$ given by Eq. S46 (for discrete frequencies ω_j), and where

$$\begin{aligned} \hat{y}_i(t) &= y_i(t) - \delta\overline{y}(t) \\ y_i(t) &\xrightarrow{\mathcal{F}} Y_i(\omega_j,T) \end{aligned} \tag{S53}$$

where $\mathcal{F}$ denotes the Fourier transform and δ can be set to 0 if the power of the target response is much smaller than the background noise or 1 if the power of the target response is expected to be comparable to the background noise.
The DOF of σ^2_{RN} (v_{RN}, the denominator of the F ratio) can be estimated by Eq. S54 for the entire epoch,

$$v_{\mathrm{RN}} = v\frac{\left[\sum_{i=1}^{N} w_i\right]^2}{\sum_{i=1}^{N} w_i^2} = vN_e \tag{S54}$$

Thus, we can now compute the probability of F by

$$p = 1 - F_{CDF}(F, v_y, v_{\mathrm{RN}}) \tag{S55}$$

Where $F_{CDF}(F, v_y, v_{\mathrm{RN}})$ is the cumulative distribution function of an F-distributed random variable with DOF v_y and v_{RN}.

Sup. 1.9.2. Fmpi in the frequency-domain—narrow-band frequency detection

There may be applications in which frequency-domain analysis is more appealing to detect the signal of interest. For example, if the stimulus evoking the response is presented periodically, such that the target response is assumed to also be a periodic signal, one may prefer to exploit this property and use a frequency-domain approach. The Fmpi, is well-suited for this purpose.
First, it is important noticing that the variance of a time-domain signal can be also estimated in the frequency domain by means of Parseval's theorem. That is,

$$\begin{aligned} \hat{\sigma}^2_y &= \frac{1}{P-1}\sum_{n=1}^{P}\left(y(t_n) - \overline{y}\right)^2 \\ &\approx \frac{1}{P(P-1)}\sum_{\omega_i>0} 2\left|Y(\omega_i,T)\right|^2 \\ y(t_n) &\xrightarrow{\mathcal{F}} Y(\omega_i,T) \text{ with } 0 \le t_n \le T \end{aligned} \tag{S56}$$

We can use this property to estimate the frequency-specific variance of the noise ($\sigma^2_{\mathrm{BN}_i}$) by summing the power spectral density components of the frequencies of interest as

$$\begin{aligned} \hat{\eta}_i(t) &= y_i(t) - \delta\overline{y}(t) \\ \hat{\eta}_i(t) &\xrightarrow{\mathcal{F}} H_i(\omega_j,T) \\ \hat{\sigma}^2_{\mathrm{BN}_i} &= \frac{1}{P(P-1)}\sum_{\omega_j} 2\left|H_i(\omega_j,T)\right|^2 \end{aligned} \tag{S57}$$

where $\omega_j \in [w_a, w_b, \ldots, w_m]$ is a set of frequencies of interest.
With that, we can estimate the frequency-specific residual noise using Eq. S17 and the F-ratio within a time window of duration T' (with P' samples) as defined in Eq. S11, in which case the frequency-specific variance of the averaged response is given by

$$\begin{aligned} \overline{y(t)} &\xrightarrow{\mathcal{F}} \overline{Y}\left(\omega_i,T'\right) \text{with } 0 \le t \le T' \le T \\ \sigma^2_{\mathrm{y}} &= \frac{1}{P'(P'-1)}\sum_{\omega_i} 2\left|\overline{Y}(\omega_j)\right|^2 \end{aligned} \tag{S58}$$

where again $\omega_j \in [w_a, w_b, \ldots, w_m]$.
The significance of the F-value depends on the DOF of σ^2_{y} and σ^2_{RN}.
Both v_y and v_{RN} can be estimated in a frequency-specific manner by Eq. S59 and Eq. S60

$$v_y = \frac{2\left[\sum_{\omega_j} \hat{G}_{\eta\eta}(\omega_j, T')\right]^2}{\sum_{\omega_j} \hat{G}_{\eta\eta}(\omega_j, T')^2} \tag{S59}$$

$$v_{\mathrm{RN}} = v\frac{\left[\sum_{i=1}^{N} w_i\right]^2}{\sum_{i=1}^{N} w_i^2} = \frac{2\left[\sum_{\omega_j} \hat{G}_{\eta\eta}(\omega_j, T)\right]^2}{\sum_{\omega_j} \hat{G}_{\eta\eta}(\omega_j, T)^2}\frac{\left[\sum_{i=1}^{N} w_i\right]^2}{\sum_{i=1}^{N} w_i^2} = \frac{2\left[\sum_{\omega_j} \hat{G}_{\eta\eta}(\omega_j, T)\right]^2}{\sum_{\omega_j} \hat{G}_{\eta\eta}(\omega_j, T)^2} N_e \tag{S60}$$

where v is similar to Eq. S59, but computed from the entire epoch of duration T, and the individuals weights w_i are the inverse of the frequency-specific variance (Eq. S57).
Note, however, that using different analysis windows for the estimation of the two variances (Eq. S57 and Eq. S58) will result in different frequency resolutions for each estimate. This can result in an imbalance between the energy contained within each frequency bin and the estimated F-value. This problem can be avoided employing the same analysis windows ($T = T'$) for both estimates.

Sup. 1.9.3. Fmpi in the frequency-domain—sparse frequency-detection

The above equations will work well if the number of samples used to estimate the background noise is large enough so that both the estimation of $\sigma^2_{\mathrm{BN}_i}$ and v_{RN} are reliable. This is usually the case when analysing a narrow-band signal with a several frequency components, e.g., the range between 100 and 2000 Hz. However, for some applications we may be interested only in few frequency bins, e.g., the fundamental frequency and several harmonics, in which case the estimation of $\sigma^2_{\mathrm{BN}_i}$ (using only a few frequency bins) will be poor.
To address this problem we can we can proceed as it follows. Consider that we have M sparsely distributed frequencies of interest $\omega_j \in [w_a, w_b, \ldots, w_m]$. Next, assume that we want to assess if the pooled power of these frequency components is significantly larger than the underlying background noise for the same frequencies. Because of the low number of frequencies to be analysed—relative to the full spectrum of the data—the estimation of $\sigma^2_{\mathrm{BN}_i}$ (Eq. S57) can be inaccurate. To avoid this problem, we can exploit the assumption that the noise at the frequencies of interest should be similar to that of their neighbouring frequencies. Under this assumption, we can obtain a good approximation of the power of the background noise at each frequency of interest by averaging the power of frequencies in their neighbourhood. That is,

$$\begin{aligned} \sigma^2_{\mathrm{BN}_i} &= \frac{1}{P(P-1)}\sum_{\omega_j} 2\overline{|Y_i(\omega_j, T)|^2} \text{ with } \omega_j > 0 \\ \overline{|Y_i(\omega_j, T)|^2} &= \frac{1}{N_{\omega_j}} \sum_{\omega_k=\omega_j-\Delta_{\omega_j}}^{\omega_j+\Delta_{\omega_j}} |Y_i(\omega_k, T)|^2 \end{aligned} \tag{S61}$$

where N_{ω_j} is the number of frequency bins around each frequency of interest used to estimate $\overline{|Y_i(\omega_j, T)|^2}$. Note that in Eq. S61 the number of samples and, therefore, the number of DOF will be higher as more samples contributed to this estimation. The number of DOF given by Eq. S60 should now include all the frequencies used to estimate $\sigma^2_{\mathrm{BN}_i}$, that is,

$$\begin{aligned} v_{\mathrm{RN}} &= v_{\mathrm{BN}}\frac{\left[\sum_{i=1}^{N} w_i\right]^2}{\sum_{i=1}^{N} w_i^2} \\ &= \frac{2\left[\sum_{\omega_n} \hat{G}_{\eta\eta}(\omega_n, T)\right]^2}{\sum_{\omega_n} \hat{G}_{\eta\eta}(\omega_n, T)^2}\frac{\left[\sum_{i=1}^{N} w_i\right]^2}{\sum_{i=1}^{N} w_i^2} \\ &= \frac{2\left[\sum_{\omega_n} \hat{G}_{\eta\eta}(\omega_n, T)\right]^2}{\sum_{\omega_n} \hat{G}_{\eta\eta}(\omega_n, T)^2} N_e \end{aligned} \tag{S62}$$

with $\omega_n \in [\omega_1 - \Delta\omega_1, \ldots, \omega_1 + \Delta\omega_1, \omega_2 - \Delta\omega_2, \ldots, \omega_2 + \Delta\omega_2, \ldots, \omega_m - \Delta\omega_m, \ldots, \omega_m + \Delta\omega_m]$.
Finally, we can estimate σ^2_y and v_y by means of equations Eq. S58 and Eq. S59 and the probability of the F-value by Eq. S55.

Sup. 1.9.4. Predicting the necessary number of epochs to detect a response

An estimation of the necessary number of epochs to detect a response can be derived by taking into account the relation between Eq. S6 and the RN as a function of the number of epochs.
An EP will be detected whenever the ongoing $F-value$ becomes greater than the critical $F-value$ ($Fc_{v_y v_{RN}}$, Eq. S63), that is, the threshold to reject the null-hypothesis.

$$F_{v_y v_{RN}} = \frac{\hat{\sigma}_y^2}{\hat{\sigma}_{RN}^2} > Fc_{v_y v'_{RN}} \tag{S63}$$

Here, the subscript v_y indicate the DOF obtained by averaging N epochs and v'_{RN} indicates the expected DOF at the target number of epochs. Next, assuming independency across epochs, we can estimate variance of the RN as a function of the number of epochs ($\tilde{\sigma}_{RN}^2(M)$) and the 'equivalent' background noise variance $\hat{\sigma}_{BN_{eq}}$ by Eq. S64

$$\begin{gathered}\tilde{\sigma}_{RN}^2(M) = \frac{\hat{\sigma}_{BN_{eq}}^2}{M} \\ \text{with } \hat{\sigma}_{BN_{eq}}^2 = N\hat{\sigma}_{RN}^2\end{gathered} \tag{S64}$$

where M is the arbitrary number of epochs.
Using Eq. S64 in Eq. S63, we can estimate the total number of necessary epochs to detect an EP, N_s, by

$$\begin{gathered}\tilde{\sigma}_{RN}^2(N_s) Fc_{v_y v'_{RN}} < \hat{\sigma}_y^2 \\ \frac{\hat{\sigma}_{BN_{eq}}^2}{max\left(\hat{\sigma}_{RN}^2, \hat{\sigma}_y^2\right)} Fc_{v_y v'_{RN}} < N_s\end{gathered} \tag{S65}$$

The denominator on the left side of Eq. S65 reflects the fact that the variance of the average response consists of the variance of the neural response (σ_x^2) and the variance of the RN, i.e., $\hat{\sigma}_y^2 = \sigma_x^2 + \hat{\sigma}_{RN}^2$, and in the absence of an EP it should not be lower than the variance of the RN. That is, N_s cannot exceed the upper limit given by

$$N_s \leq \frac{\hat{\sigma}_{BN_{eq}}^2}{\hat{\sigma}_{RN}^2} Fc_{v_y v'_{RN}} = NFc_{v_y v'_{RN}} \tag{S66}$$

The DOF v_y is given by Eq. S52 or Eq. S59, and v'_{RN} is derived from the current v_{RN} projected at N_s as

$$v'_{RN} = \frac{v_{RN}}{N_e}(N_e + N_s - N) \text{ where } N_s \in A = [N, N+1, N+2, N+3, \ldots] \tag{S67}$$

Here, N_s is $min(A)$ for which Eq. S65 satisfied.

Sup. 1.10. Analytical DOF estimation

Consider a noise with a PSD that follows a $\frac{1}{f^p}$ function, with $f > 0$ being the frequency. The PSD of the noise passed through a linear filter—Y_{yy}—is given by

$$Y_{yy}(f) = |H(f)|^2 G_{yy}(f) = |H(f)|^2 \left|\frac{1}{f^p}\right| \tag{S68}$$

Now, consider an ideal bandpass filter (Eq. S69) and the continuous expression of the DOF given by Eq. S51

$$H(f) = \begin{cases} 1 & f_l < f < f_h \\ 0 & \text{otherwise} \end{cases} \tag{S69}$$

We can obtain the analytical DOF for bandpass-filtered $1/f^p$ PSD noise within a given analysis time window of duration T as

$$\begin{gathered} v = \frac{2\left[\int_0^\infty G_{yy}(f)\,df\right]^2}{\frac{1}{T}\int_0^\infty G_{xx}(f)^2\,df} = \frac{2\left[\int_{f_l}^{f_h}\left|\frac{1}{f^p}\right|df\right]^2}{\frac{1}{T}\int_{f_l}^{f_h}\left|\frac{1}{f^p}\right|^2 df} = \frac{a}{b} \\ a = \begin{cases} 2\left[\frac{f_h^{1-p} - f_l^{1-p}}{1-p}\right]^2 & p \neq 1 \\ 2log(\frac{f_h}{f_l})^2 & p = 1 \end{cases} \\ b = \begin{cases} \frac{1}{T}\frac{f_h^{1-2p} - f_l^{1-2p}}{1-2p} & p \neq \frac{1}{2} \\ \frac{1}{T}log(\frac{f_h}{f_l}) & p = \frac{1}{2} \end{cases} \end{gathered} \tag{S70}$$

SUPPLEMENTARY FIGURES

[Sup. Figure 1 about here.]

[Sup. Figure 2 about here.]

[Sup. Figure 3 about here.]

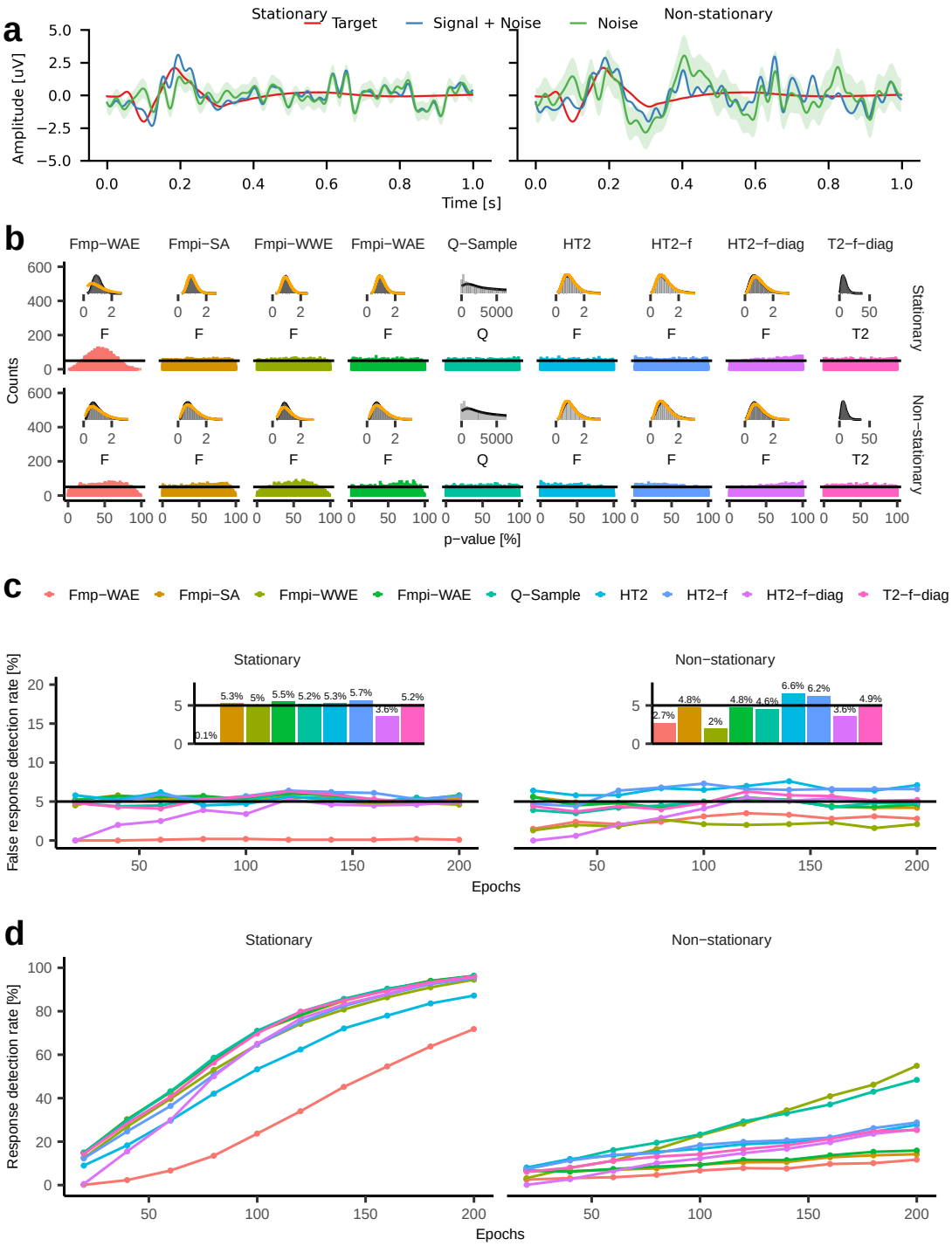


Figure 1. Fmpi sensitivity and specificity are robust across time under stationary and non-stationary noise in simulated cortical responses. a Examples of simulations under stationary (left) and non-stationary (right) background noise. The template cortical response (Target) is shown by red lines. Weighted average waveforms containing the cortical template are shown by light blue lines (Signal + Noise), whilst averages obtained without cortical responses (Noise) are shown in green, with the standard error of the mean indicated by the shaded green area. **b** P-value histograms (200 bins) obtained in the absence of cortical responses for stationary (top) and non-stationary (bottom) noise. The horizontal solid black lines represent the ideal unbiased distribution. Insets show the empirical distribution of the F-, Q-, or T2-values versus the theoretical (orange line) probability density functions (PDFs) of an F-distribution pooled across epochs. The DOF, v_1 and v_2, correspond to the median of the pooled DOF estimated for each method at each epoch block. **c** False response detection rates (noise-only data) obtained at a nominal sensitivity of 95%. Insets show false response detection rates pooled across all epochs. **d** response detection rates in the presence of a cortical response (Signal + Noise) as a function of the number of epochs. A total of 1000 simulations were employed in each epoch. Left panels show stationary noise; right panels show non-stationary noise. The number of simulations per epoch block is indicated by the lines connected to the data points.

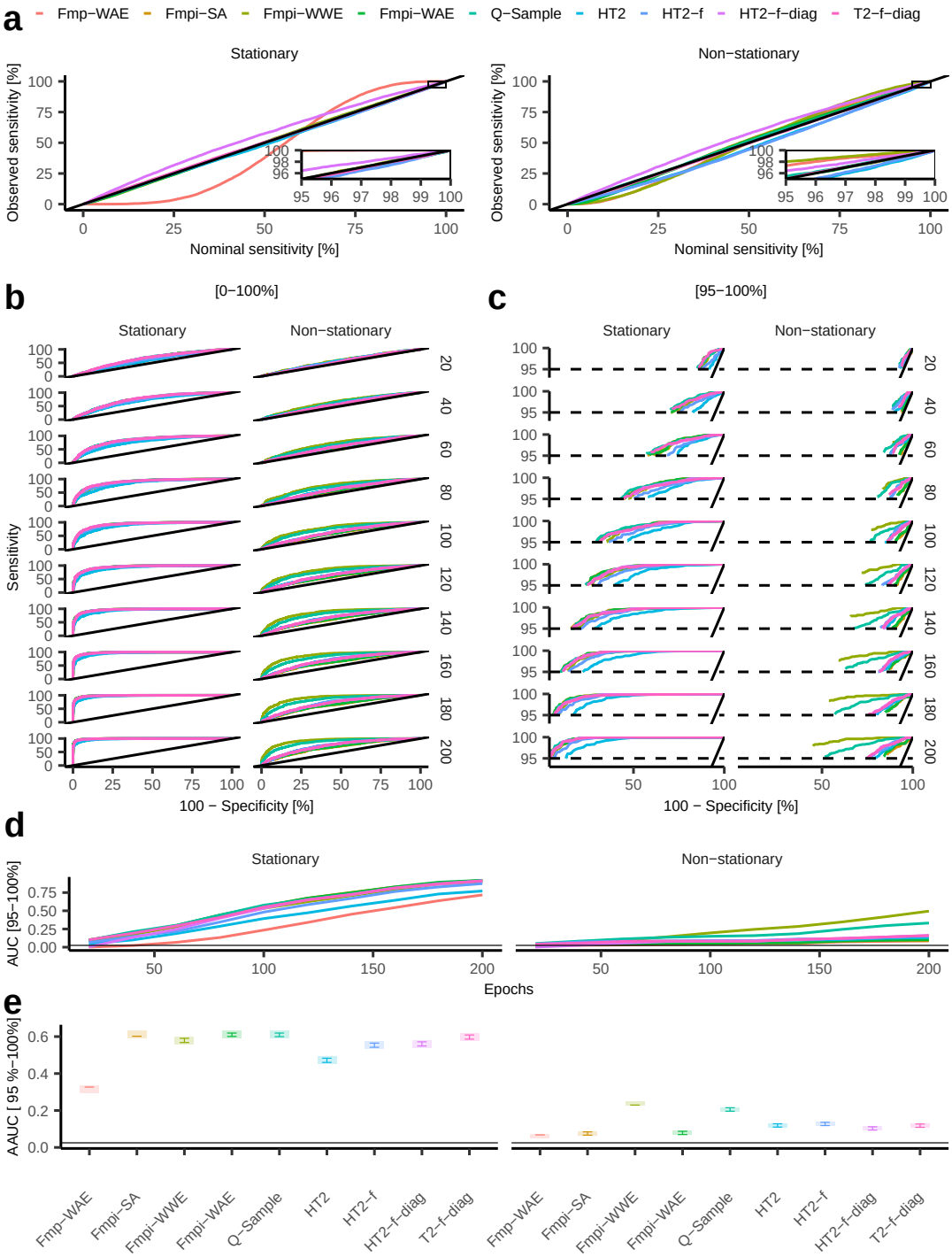


Figure 2. Fmpi demonstrates high performance across the span of clinically relevant sensitivity rates in simulated cortical responses. a Nominal versus observed sensitivity for stationary (left) and non-stationary (right) noise across all possible sensitivities. Traces represent data pooled across all epochs. The diagonal solid black line indicates the ideal unbiased expectation. Insets provide a zoomed view of the clinically relevant sensitivity regions (95-100%). **b** ROCs for stationary (left) and non-stationary (right) noise for all possible sensitivities at different number of averaged epochs (indicated on the right side of the panels). The diagonal solid black line represent the chance level of the ROCs. **c** Same as **b**, but computed for the clinically relevant nominal sensitivities (95-100%). The 95% nominal sensitivity is indicated by horizontal dashed lines. **d** AUC as a function of the number of epochs for the clinically relevant nominal sensitivities shown in **c**. Note that the AUCs have been normalised by the total area within the clinically relevant region. The corresponding chance level is shown by the horizontal solid black line. **e** AAUC for the clinically relevant nominal sensitivities. Shaded areas represent 95% confidence intervals. The length of the error bars has been adjusted to allow for visual post hoc comparisons (Holm FWER corrected). Non-overlapping segments indicate significant differences, whilst overlapping segments indicate non-significantly differences. The horizontal solid black line indicates the chance level of the AAUC.

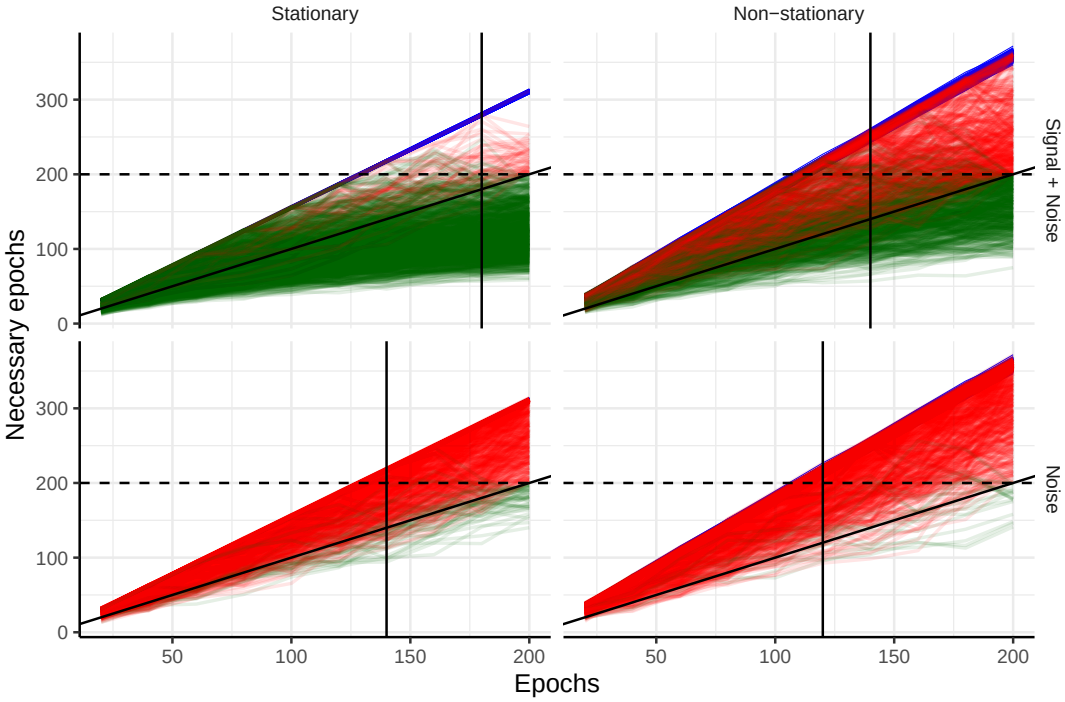


Figure 3. Fmpi enables reliable prediction of futility in simulated cortical recordings. The horizontal axes indicates the ongoing number of recorded epochs, whilst the vertical axes show the estimated number of epochs required to detect a significant cortical response. Results are based on 1000 simulations (each line corresponds to an individual simulation from Figure 1). Top panels show cortical simulations with Signal + Noise; bottom panels show, noise only data. Columns represent stationary and non-stationary noise. Green traces depict the trajectory of the estimated epochs required to detect a significant response, where detection occurred at the end of the simulation (at 200 epochs). Red traces indicate simulations in which a cortical response was not detected. Blue traces show the upper limit of the estimation for each simulated recording, determined by the absence of a neural response ($\hat{\sigma}_y^2 = \hat{\sigma}_{RN}^2$, right side of Eq. 18). The diagonal black solid line marks the detection boundary: if the predicted number of epochs falls below the diagonal, the response has been detected; if it lies between the region determined by the noise (blue) and the diagonal, the response has not been detected. When the predicted number of epochs crosses the futility limit (horizontal black dashed lines), waiting for a response is considered futile. This estimation assumes that the overall recording conditions remain unchanged. Vertical black solid lines indicate the median number of epochs at which the futility threshold was exceeded. Traces were obtained using the Fmpi-WWE approach at a sensitivity of 95%.

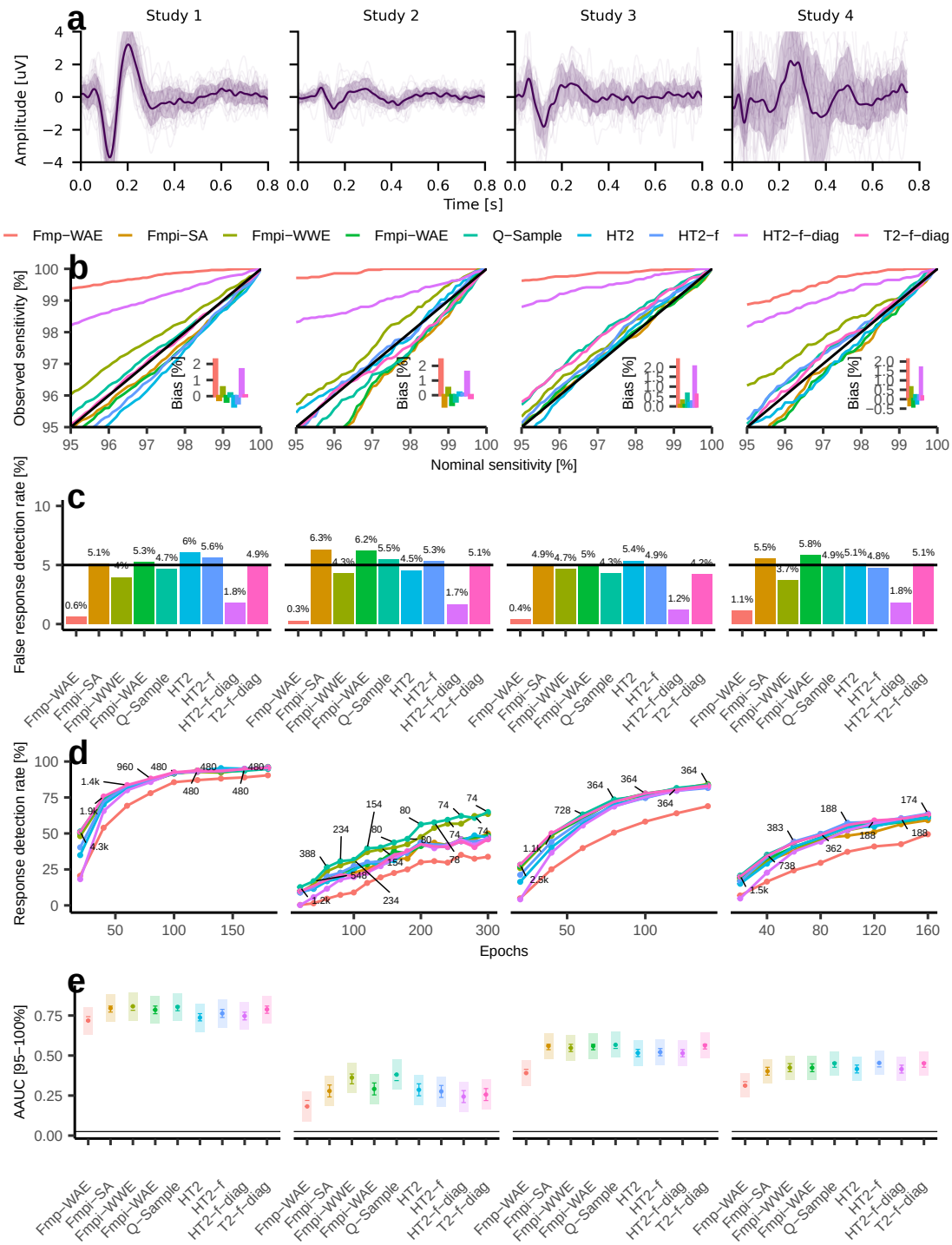


Figure 4. Fmpi demonstrates robust performance in the detection of human cortical responses. a Example final averaged cortical EPs (bold lines) with individual data (thin lines) and standard deviation (shaded area) from different studies (each column represents a different study). **b** Nominal versus observed sensitivity for clinically relevant nominal sensitivity thresholds (95-100%). Traces represent data pooled across all epochs. The diagonal solid black line indicates the ideal unbiased expectation. Insets show the average bias across nominal sensitivities. A positive bias indicates that the obtained sensitivity was higher than the nominal value, whilst a negative bias indicates the opposite. **c** False response detection rates (derived from the same data) obtained for a nominal sensitivity of 95% pooled across all epochs. **d** Response detection rates for each study as a function of the number of epochs. The number of independent measurements contributing to each epoch is indicated by the numbers connected to the data points. **e** AAUC for the clinically relevant nominal sensitivities. Shaded areas represent 95% confidence intervals. The length of the error bars has been adjusted for visual post hoc comparisons (Holm FWER corrected). Non-overlapping segments indicate significant differences, whilst overlapping segments indicate non-significant differences. The horizontal solid black line indicates the chance level of the AAUC. The different methods are indicated by the colours across all panels.

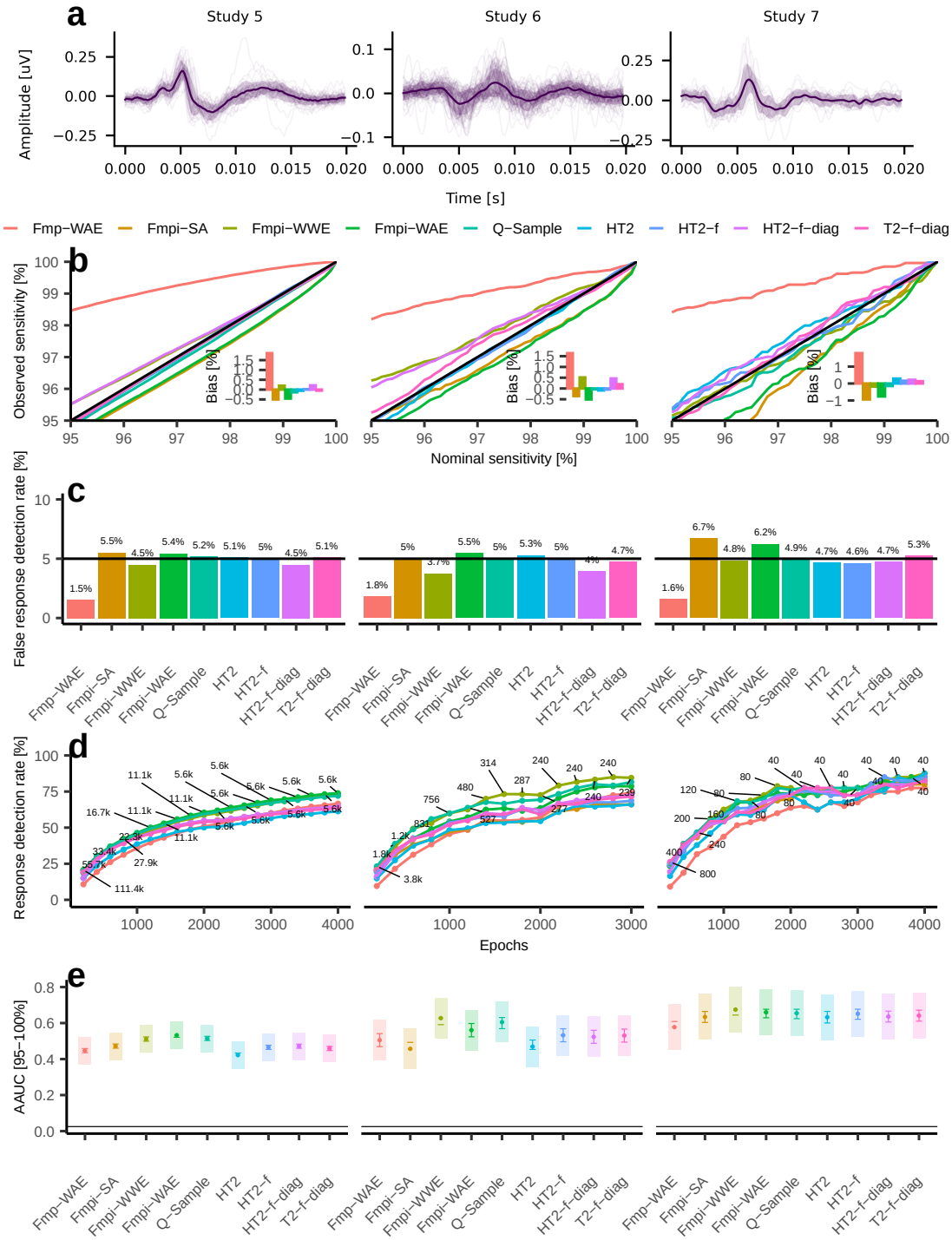


Figure 5. Fmpi demonstrates robust performance in the detection of human ABRs. a Example averaged ABRs (bold lines) with individual data (thin lines) and standard deviation (shaded area) from different studies (each column represents to a different study). **b** Nominal versus observed sensitivity for the clinically relevant nominal sensitivities (95-100%). Traces represent data pooled across all epochs. The diagonal solid black line indicates the ideal unbiased expectation. Insets show the average bias across nominal sensitivities. A positive bias indicates that the obtained sensitivity was higher than the nominal value, whilst a negative bias indicates the opposite. **c** False response detection rates (derived from the same data) obtained for a nominal sensitivity of 95% pooled across all epochs. **d** Response detection rates for each study as a function of the number of epochs. The number of independent measurements contributing to each epoch is indicated by the numbers connected to the data points. **e** AAUC for the clinically relevant nominal sensitivities. Shaded areas correspond to 95% confidence intervals. The length of the error bars has been adjusted for visual post hoc comparisons (Holm FWER corrected). Non-overlapping segments indicate significant differences, whilst overlapping segments indicate non-significantly differences. The horizontal solid black line indicates the chance level of the AAUC. The different methods are indicated by the colours across all panels.

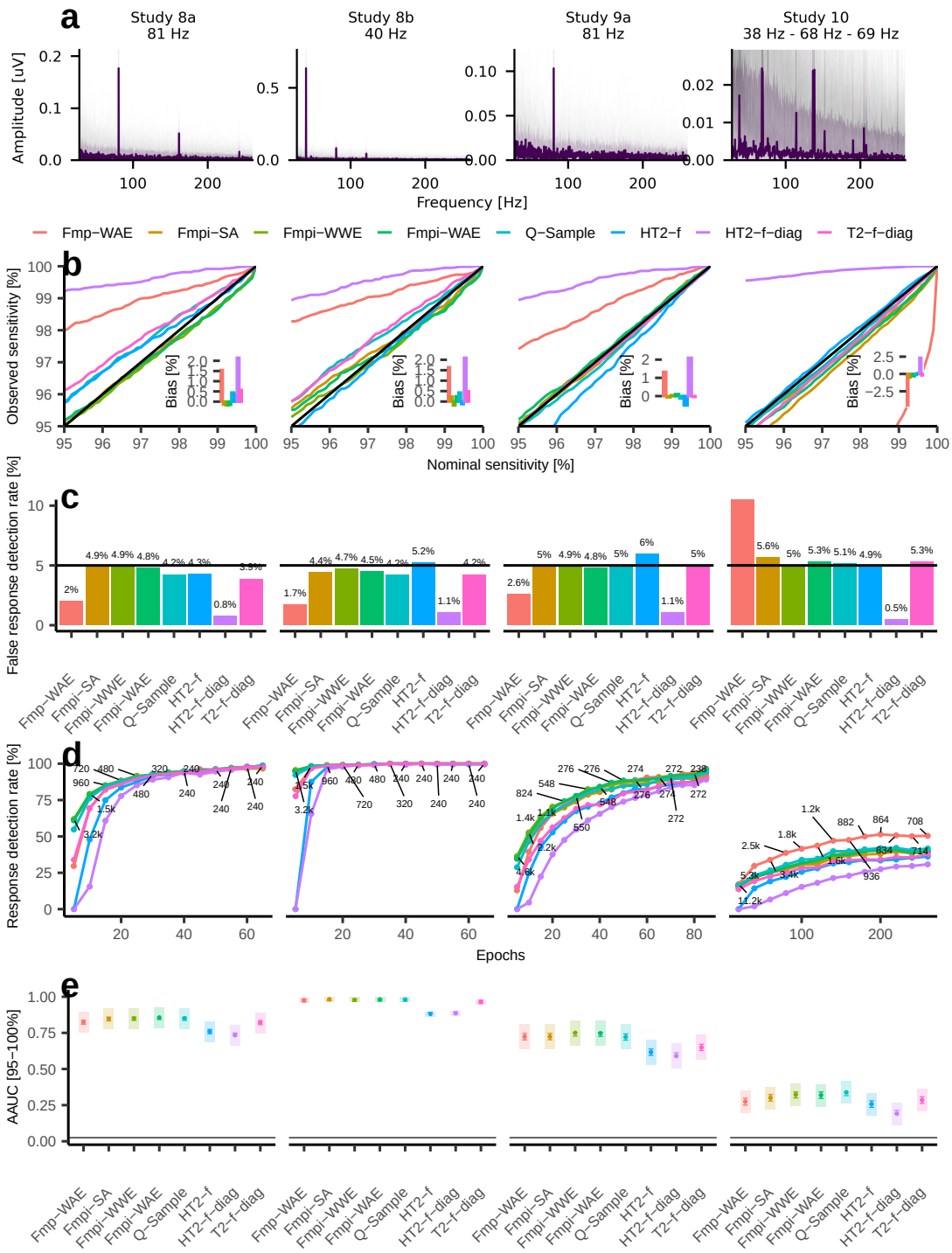


Figure 6. Fmpi demonstrates robust performance in the detection of human ASSRs at high modulation rates. a Example ASSRs from different studies (each column represents a different study). Responses were obtained for ASSR rates of ≈ 40 and 81 Hz (study 8), 81 Hz (study 9), and 38, 68 and 69 Hz, simultaneously (study 10). Three harmonics were included in studies 8 and 9, whilst 12 harmonics were included in study 10. **b** Nominal versus observed sensitivity for the clinically relevant nominal sensitivities (95-100%). Traces represent data pooled across all epochs. The diagonal solid black line indicates the ideal unbiased expectation. Insets show the average bias across nominal sensitivities. A positive bias indicates that the obtained sensitivity was higher than the nominal value, whilst a negative bias indicates the opposite. **c** False response detection rates (derived from the same data) obtained for a nominal sensitivity of 95% pooled across all epochs. **d** Response detection rates for each study as a function of the number of epochs. The number of independent measurements contributing to each epoch is indicated by the numbers connected to the data points. **e** AAUC for the clinically relevant nominal sensitivities. Shaded areas correspond to 95% confidence intervals. The length of the error bars has been adjusted for visual post hoc comparisons (Holm FWER corrected). Non-overlapping segments indicate significant differences, whilst overlapping segments indicate non-significantly differences. The horizontal solid black line indicates the chance level of the AAUC. The different methods are indicated by the colours across all panels.

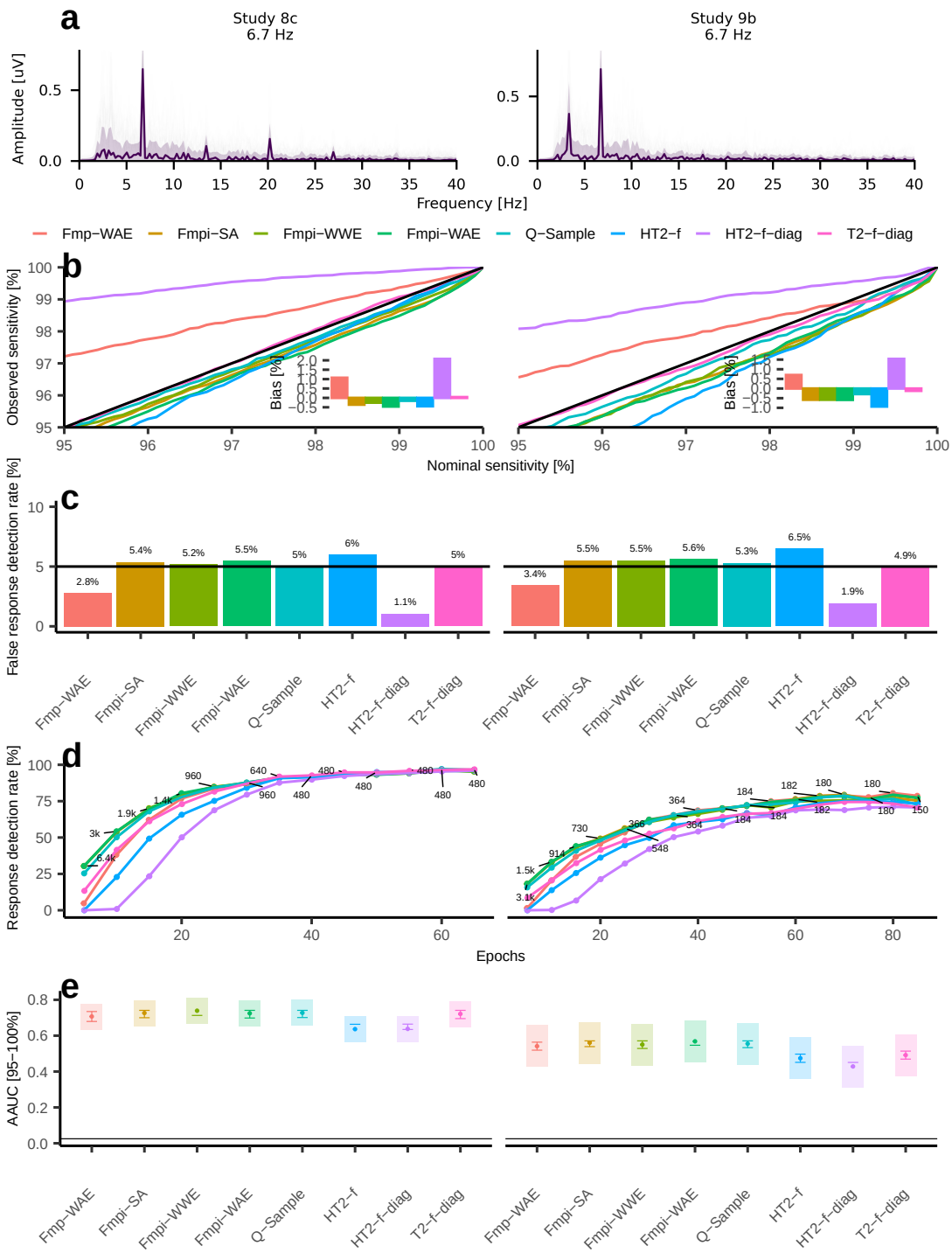


Figure 7. Fmpi demonstrates robust performance for the detection of human ASSRs at low modulation rates. a Example averaged ASSRs (bold lines) with individual data (thin lines) and standard deviation (shaded area) from different studies (each column represents a different study). Responses were obtained for ASSR rates of ≈ 7 Hz and the next two harmonics. **b** Nominal versus observed sensitivity for the clinically relevant nominal sensitivities (95-100%). Traces represent data pooled across all epochs. The diagonal solid black line indicates the ideal unbiased expectation. Insets show the average bias across nominal sensitivities. A positive bias indicates that the obtained sensitivity was higher than the nominal value, whilst a negative bias indicates the opposite. **c** False response detection rates (derived from the same data) obtained for a nominal sensitivity of 95% pooled across all epochs. **d** Response detection rates for each study as a function of the number of epochs. The number of independent measurements contributing to each epoch is indicated by the numbers connected to the data points. **e** AAUC for the clinically relevant nominal sensitivities. Shaded areas correspond to 95% confidence intervals. The length of the error bars has been adjusted for visual post hoc comparisons (Holm FWER corrected). Non-overlapping segments indicate significant differences, whilst overlapping segments indicate non-significantly differences. The horizontal solid black line indicates the chance level of the AAUC. The different methods are indicated by the colours across all panels.

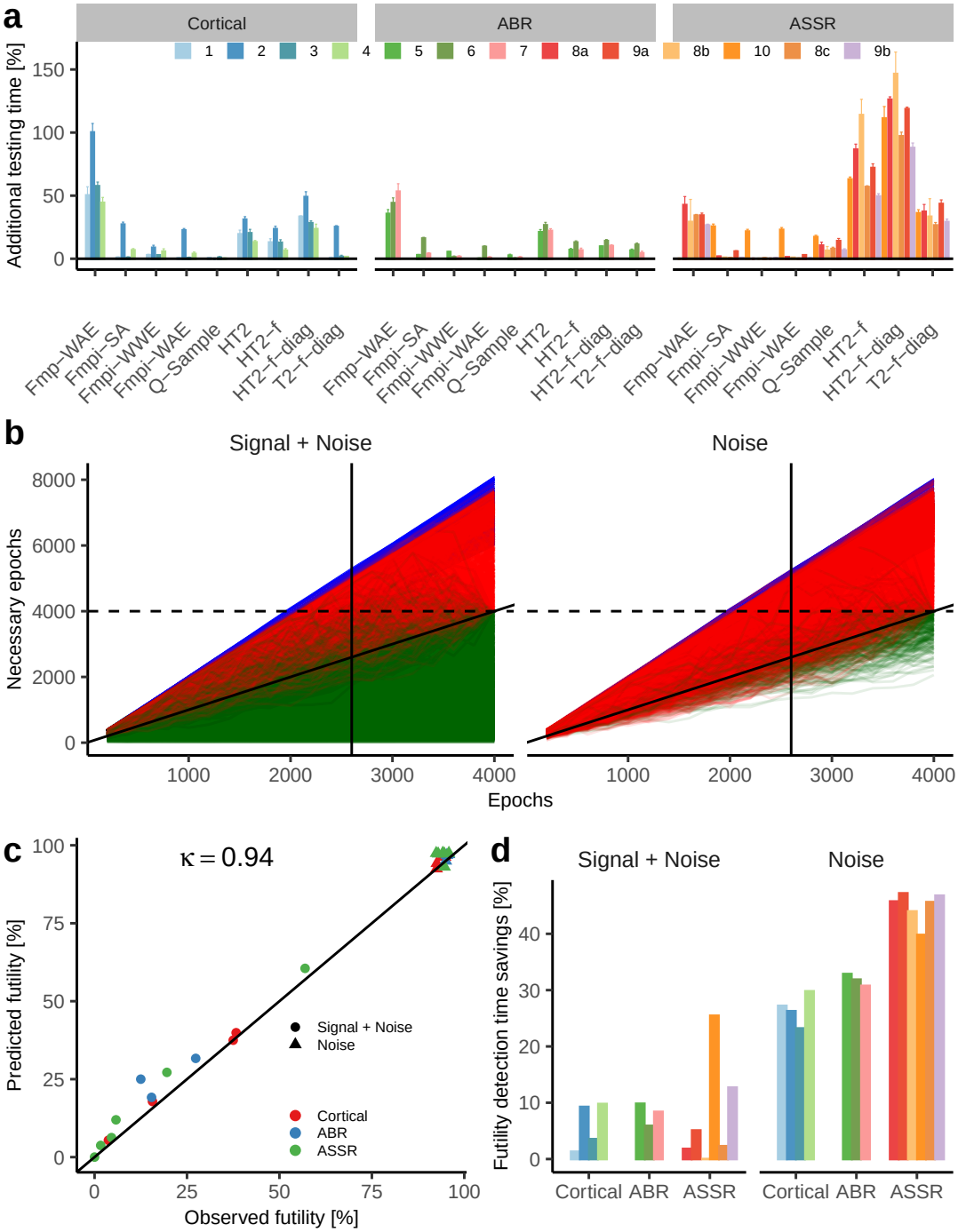


Figure 8. The Fmpi detector substantially reduces the testing time required for EP detection. a Normalised testing times for each detection method grouped by EP type (Cortical, ABR, and ASSR). Normalised times are averaged across all clinically relevant nominal sensitivities (95-100%). Study identifiers are indicated by colour. **b** Futility curves for ABRs from study 5 (n = 5570 traces). The left panel shows results for original data (Signal + Noise), whilst the right panel present results for no-response data derived from the same dataset (Noise). Green traces depict the trajectory of the ongoing (horizontal axis) versus the predicted (vertical axis) number of epochs for each measurement in which a response was detected at the end of the recording (4000 epochs). Red traces indicate measurements where the futility threshold was exceeded (horizontal dashed line). Blue traces show the upper limit of the estimation (determined by the noise; right side of Eq. 18) for each recording. The diagonal solid black line denotes the detection boundary. A trace falling below the diagonal indicates EP detection. Vertical black solid lines indicate the median number of epochs at which futility was detected. **c** Relationship between observed (horizontal axis) and predicted (vertical axis) futility rates across all studies. **d** Futility-based time savings across all studies, shown for both the presence (Signal + Noise) and absence (Noise) of an EP.

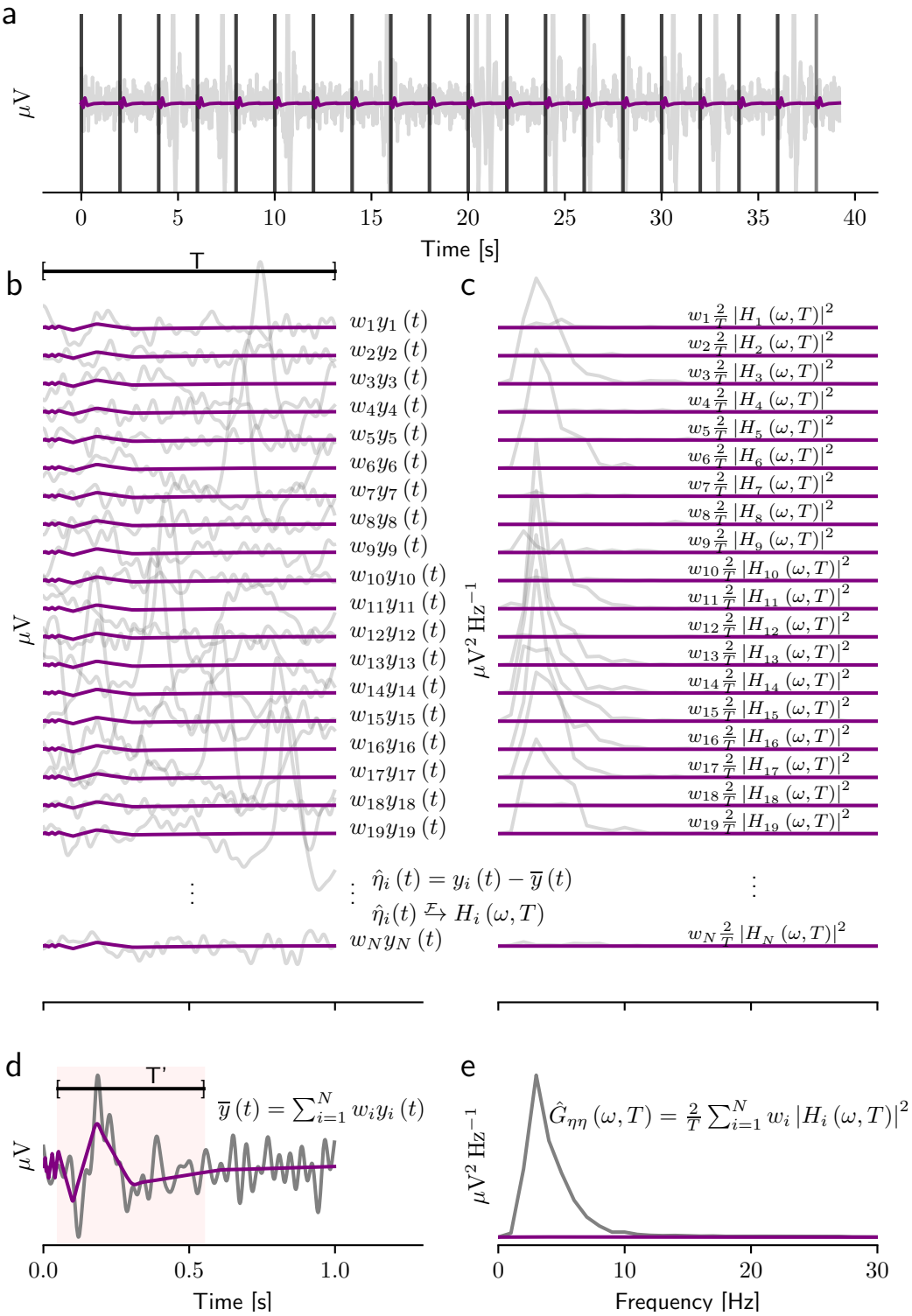


Figure 9. Evoked response estimation. The target EP is shown in purple. **a** Continuous time series consisting of the EPs and the recording BN is shown by the grey line. The timing of the event evoking the response is shown by the vertical black lines. **b** Single epochs (or epochs) of duration T $(y_1(t), y_2(t), \ldots y_N(t))$ starting at the time of each event. Each of these epochs is assigned a weight $(w_1, w_2, \ldots w_N)$. **c** Power spectral density (PSD)—$\frac{2}{T}|H_i(\omega, T)|^2$—of the estimated noise $(\eta_i(t))$ at each epoch. **d** Weighted averaged response in the time domain, and **e** weighted average PSD of the noise in the frequency domain.

Table 1. Summary of the data and EPs analysed in this study. Columns from left to right: study ID; data source; EP type; stimulus; number of participants (N); total number of measurements (M); number of EEG channels included (C); age, hearing status (normal hearing (NH) or hearing-impaired (HI)) ; ethical approval; and EEG recording system. The studies encompass cortical ACC responses to interaural phase differences (IPDs), interaural time differences (ITDs), frequency, or spectrotemporal transitions; cortical auditory evoked potentials (AEPs) to speech sounds; ASSRs to various amplitude modulations (AMs) and binaural stimuli (interaural phase modulation following responses (IPM-FRs) and interaural time difference following responses (ITD-FRs); and ABRs to narrow-band CE-chirp, tone-bursts, and clicks.

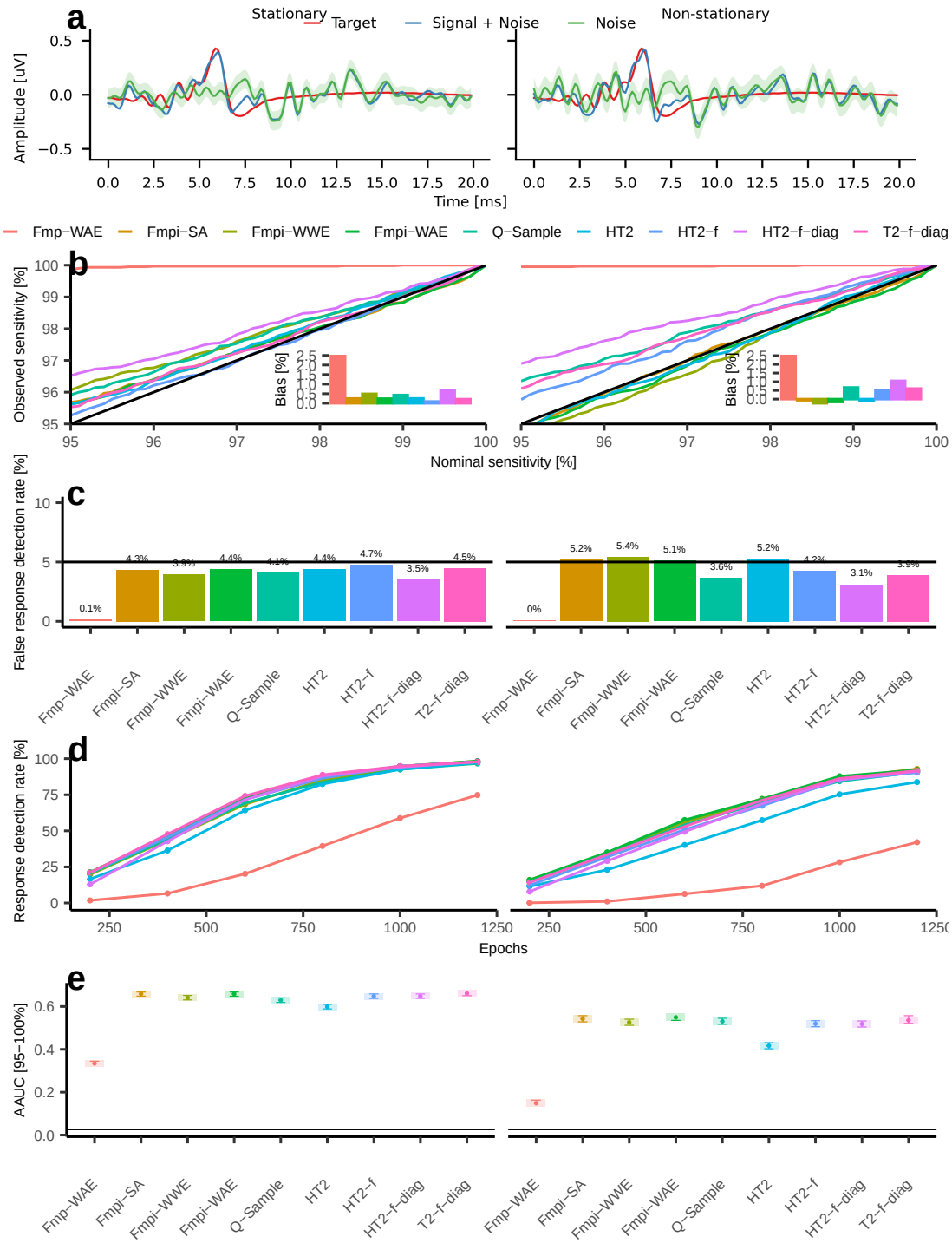
Study	Source	Type	Stimulus	N	M	C	Age	Hearing	Ethics	EEG system
1	Van Yper et al.[89]	Cortical ACC	IPM; -90/90° & 0/180° IPD changes; 40 and 80-Hz AMs; test-retest	30	240	2	18-42 years	NH	Macquarie University ethics committee (52021646228840)	BioSemi ActiveTwo (Amsterdam, The Netherlands)
2	Simonsen et al.[90]	Cortical ACC	Spectrotemporal modulations; 0-to-16 dB contrast levels	16	40	2	20-79 years	NH & HI (audibility compensated)	Science-Ethics Committee for the Capital Region of Denmark (H-1-2013-138)	Modified Eclipse system (Interacoustics, Middelfart, Denmark)
3	Rekswinkel et al.[91]	Cortical ACC	Spectral changes; vowel-like and pure tones stimuti[7]; ACC responses to frequency and/or amplitude changes	45	182	2	21-81 years	NH & HI (audibility compensated)	Science-Ethics Committee for the Capital Region of Denmark (H-1-2013-138)	BioSemi ActiveTwo (Amsterdam, The Netherlands)
4	Chesnaye et al.[44,92]	Cortical AEP	Synthetic speech tokens[93]; mid and high frequency; 0 to 30 dB SL	48	188	1	3-7 months	HI (aided)	North West National Research Ethics Service Ethics Committee (15/NW/0736)	Modified Eclipse system (Interacoustics, Middelfart, Denmark)
5	Kristensen et al.[94]	ABR	NB CE-Chirp[95]; left & right ears; 0.5, 1, 2, & 4 kHz; 35-85 dB nHL; consecutive or simultaneous presentation; test & retest	25	2785	2	20-29 years	NH	Science-Ethics Committee for the Capital Region of Denmark (H-1-2013-138)	Modified Eclipse system (Interacoustics, Middelfart, Denmark)
6	Lightfoot & Stevens[96]	ABR	4kHz tone burst; 30 to 85 dB nHL	25	240	1	1-12 weeks	Referred from the English newborn screen	Arrowe Park Hospital, Wirral, United Kingdom	Modified Eclipse system (Interacoustics, Middelfart, Denmark)
7	Bowen et al.[97]	ABR	100 μs clicks; 60 peSPL; diotic stimulation	20	20	2	19-42 years	NH	Macquarie University ethics committee (5201600895)	BioSemi ActiveTwo (Amsterdam, The Netherlands)
8a	Van Yper et al.[89]	ASSR	81-Hz AMs under -90/90° & 0/180° IPD changes; test-retest	30	120	2	18-42 years	NH	Macquarie University ethics committee (52021646228840)	BioSemi ActiveTwo (Amsterdam, The Netherlands)
8b	Van Yper et al.[89]	ASSR	40-Hz AMs under -90/90° & 0/180° IPD changes; test-retest	30	120	2	18-42 years	NH	Macquarie University ethics committee (52021646228840)	BioSemi ActiveTwo (Amsterdam, The Netherlands)
8c	Van Yper et al.[89]	ASSR	6.7-Hz IPM-FRs; -90/90° & 0/180° IPD transitions; 40 and 81-Hz AMs; test & retest	30	240	2	18-42 years	NH	Macquarie University ethics committee (52021646228840)	BioSemi ActiveTwo (Amsterdam, The Netherlands)
9a	Undurraga et al.[98]	ASSR	81-Hz AMs under 6 ITDs	23	138	2	18-35 years	NH	Macquarie University ethics committee (52020527017402)	BioSemi ActiveTwo (Amsterdam, The Netherlands)
9b	Undurraga et al.[98]	ASSR	6.7-Hz ITD-FRs; 4 ITDs; 81-Hz AM	23	92	2	18-35 years	NH	Macquarie University ethics committee (52020527017402)	BioSemi ActiveTwo (Amsterdam, The Netherlands)
10	Visram et al. (in preparation)	ASSR	38.1, 68.4, and 69.3-Hz simultaneous AMs for 0.7, 2, and 4-kHz carriers, respectively, using speech-like NB CE-Chirp[99]	40	375	2	3-7 months	HI (aided)	North West National Research Ethics Service Ethics Committee (15/NW/0736).	Modified Eclipse system (Interacoustics, Middelfart, Denmark)

Table 2. Additional detection times. Columns, from left to right: study ID; EP type; and the various detection methods. The additional detection time (in %) was estimated as the ratio of the testing time for each detector to the shortest detection time, calculated for nominal sensitivities between 95 and 100% (in increments of 0.01%, excluding 100%). The standard deviation is shown in parentheses (in %).

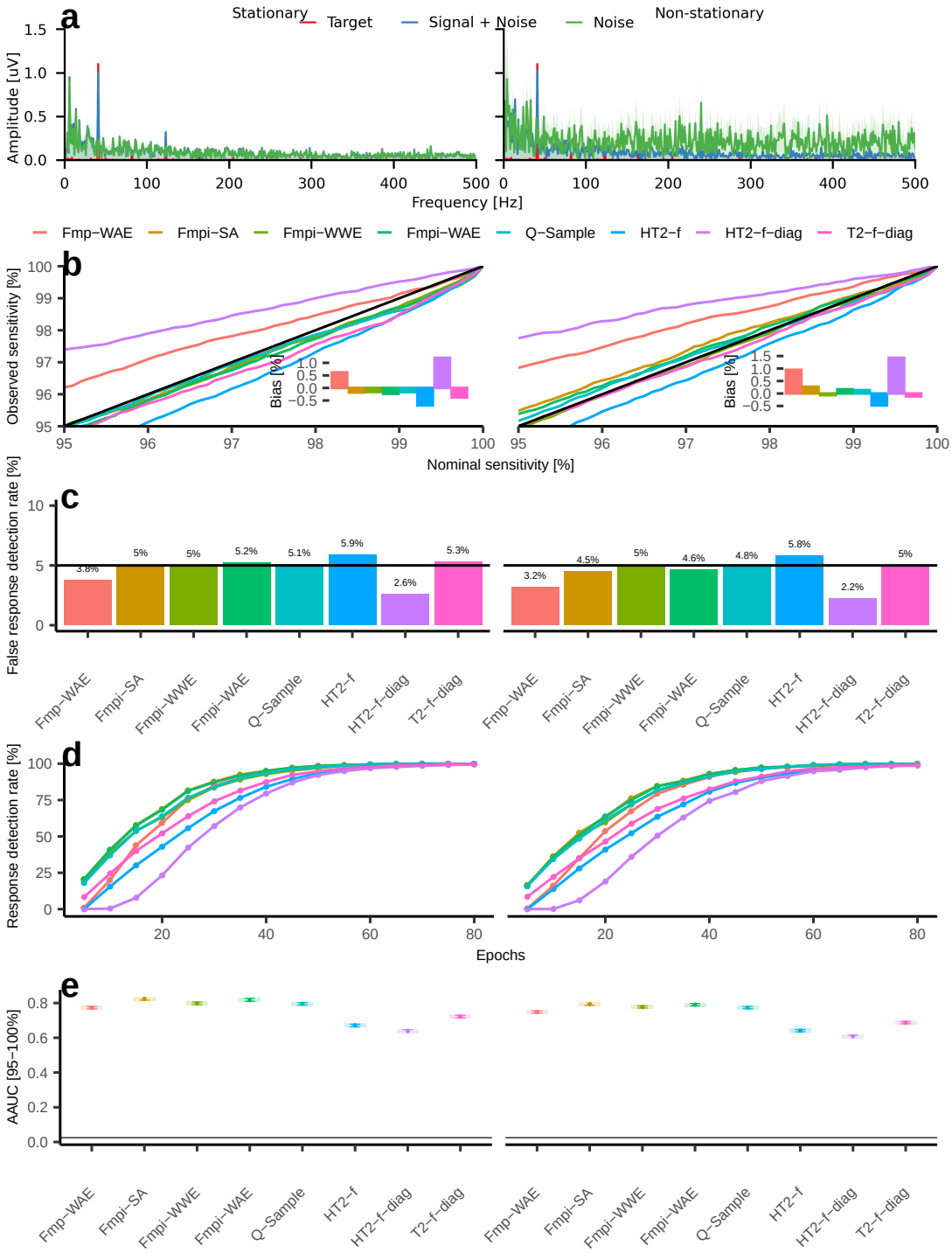
Study	EP	Fmp-WAE	Fmpi-SA	Fmpi-WWE	Fmpi-WAE	Q-Sample	HT2	HT2-f	HT2-f-diag	T2-f-diag
1	Cortical	50.2 (6.9)	0.7 (0.7)	2.7 (0.8)	0.4 (0.7)	0.5 (0.5)	19.4 (3.3)	12.9 (3.0)	33.2 (0.7)	0.4 (0.8)
2	Cortical	100.1 (7.1)	27 (2.0)	8.7 (2.0)	22.4 (1.5)	0 (0.0)	30.9 (2.4)	23.4 (2.3)	48.9 (4.2)	25.1 (1.1)
3	Cortical	57.5 (3.3)	0.7 (0.8)	2.7 (0.6)	0.5 (0.8)	0.8 (0.9)	20.2 (3.1)	12.6 (2.4)	28.2 (1.9)	1.4 (1.3)
4	Cortical	44.3 (4.3)	6.6 (1.4)	5.6 (2.2)	3.8 (1.5)	0 (0.2)	13 (1.5)	6.3 (1.8)	23.5 (4.0)	1.3 (0.6)
5	ABR	35.5 (3.5)	2.8 (0.6)	5.3 (0.4)	0 (0.4)	2.5 (1.1)	21 (2.1)	6.9 (1.2)	9.6 (0.4)	6.4 (1.3)
6	ABR	44.1 (4.2)	16 (1.0)	1.2 (1.0)	9.3 (0.8)	0.1 (0.4)	26.5 (2.3)	12.8 (1.3)	14 (1.2)	11.1 (1.4)
7	ABR	53.1 (6.3)	3.8 (0.9)	1.4 (0.9)	0.7 (1.0)	0.8 (1.0)	22 (1.8)	6.5 (1.9)	10 (1.0)	4.3 (1.7)
8a	ASSR	42.5 (6.7)	1.7 (0.3)	0 (0.0)	1.2 (0.2)	10.4 (2.7)	NA	86.5 (4.4)	126 (2.3)	37.2 (5.8)
9a	ASSR	34.2 (1.9)	5.7 (0.7)	0 (0.0)	2.7 (0.3)	14 (2.1)	NA	71.9 (3.3)	118.5 (1.6)	43.4 (3.3)
8b	ASSR	29.1 (17.9)	0.8 (0.2)	0 (0.0)	0.8 (0.2)	6.3 (3.4)	NA	113.8 (12.5)	146.4 (17.4)	33.3 (14.3)
10	ASSR	0 (0.0)	25.4 (2.1)	21.5 (1.8)	22.9 (1.9)	17.2 (1.4)	NA	62.8 (1.9)	111.2 (9.4)	36.1 (2.8)
8c	ASSR	34.1 (1.0)	0.4 (0.4)	0.2 (0.4)	0 (0.3)	7.7 (1.5)	NA	56.8 (1.0)	97 (3.3)	26.4 (2.3)
9b	ASSR	26.3 (1.1)	0.6 (0.7)	0.4 (0.6)	0.2 (0.6)	6.5 (1.4)	NA	49.3 (2.2)	87.7 (3.9)	29 (2.4)

Table 3. Futility detection across studies. Columns, from left to right: study ID; EP type; type of data analysed (original recording: Signal + Noise; or derived noise data: Noise); total number of measurements analysed (N); futility threshold (in epochs); number of detected EPs; expected futility (%) based on the total number of detected responses; detected futility (%); difference between expected and detected futility (%); agreement (%) between expected and observed futillity based on individual measurements; and total savings (as a percentage of epochs) based on measurements being terminated when futility was triggered (%). Analyses were conducted at a nominal sensitivity of 95% employing the Fmpi-WWE detector.

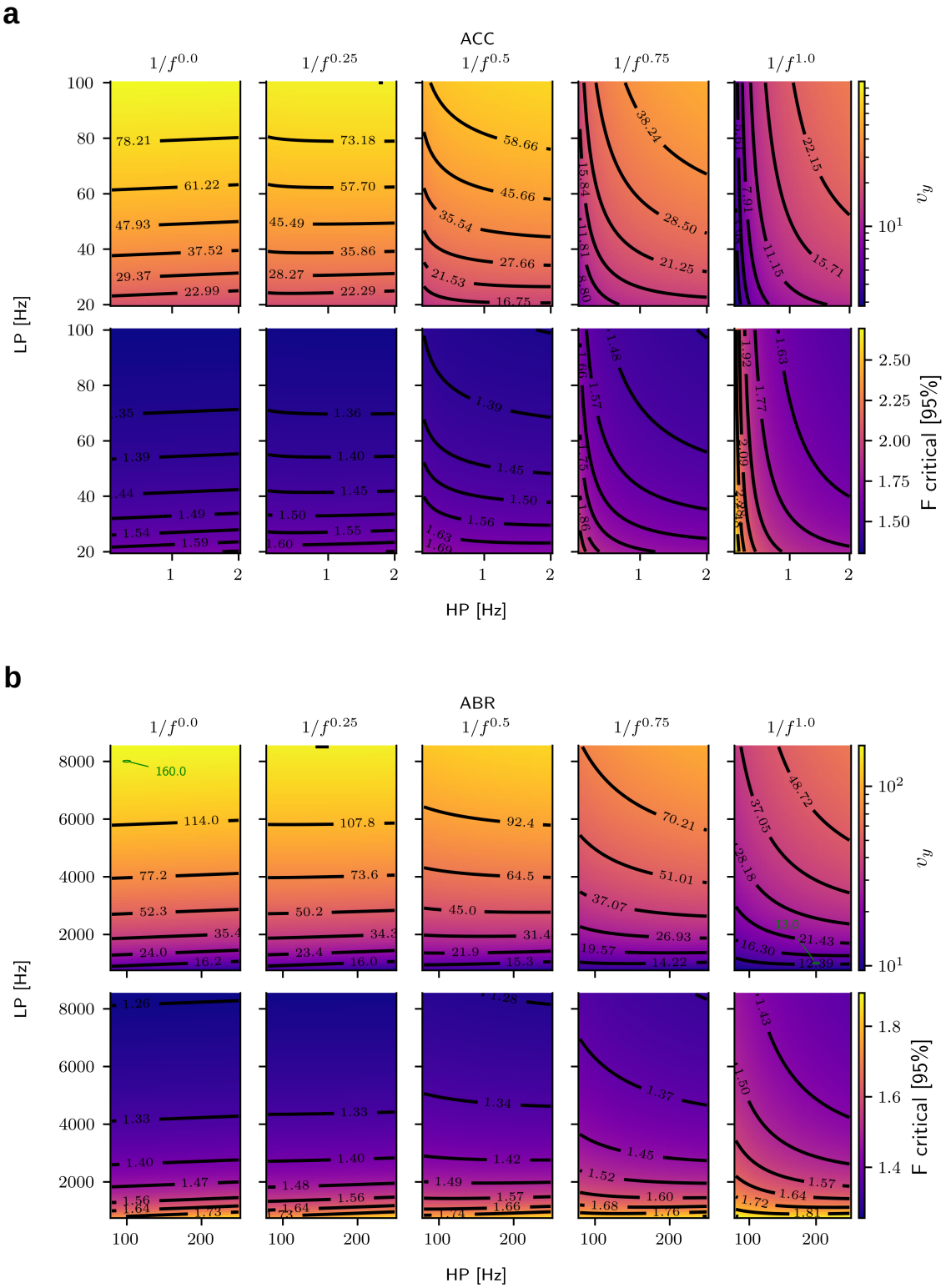
Study ID	EP	Source	N	Futility threshold	Detected responses	Expected futility [%]	Detected futility [%]	Difference [%]	Agreement [%]	Total Savings [%]
1	Cortical	Signal + Noise	480	180	462	3.8	5.4	1.6	98.3	1.3
1	Cortical	Noise	480	180	25	94.8	96.2	1.4	98.5	27.2
2	Cortical	Signal + Noise	80	300	50	37.5	37.5	0.0	96.2	9.3
2	Cortical	Noise	80	300	6	92.5	92.5	0.0	98.8	26.2
3	Cortical	Signal + Noise	364	140	307	15.7	17.9	2.2	97.8	3.5
3	Cortical	Noise	364	140	27	92.6	94.2	1.6	98.4	23.2
4	Cortical	Signal + Noise	188	160	116	38.3	39.9	1.6	98.4	9.8
4	Cortical	Noise	188	160	11	94.1	96.3	2.2	97.9	29.8
5	ABR	Signal + Noise	5570	4000	4046	27.4	31.7	4.3	95.7	9.8
5	ABR	Noise	5570	4000	262	95.3	96.8	1.5	98.5	32.8
6	ABR	Signal + Noise	240	3000	203	15.4	19.2	3.8	96.2	5.9
6	ABR	Noise	240	3000	10	95.8	97.1	1.3	98.8	31.9
7	ABR	Signal + Noise	40	4000	35	12.5	25.0	12.5	87.5	8.4
7	ABR	Noise	40	4000	2	95.0	95.0	0.0	100.0	30.8
8a	ASSR	Signal + Noise	240	65	236	1.7	3.8	2.1	97.9	1.8
8a	ASSR	Noise	240	65	10	95.8	97.5	1.7	98.3	45.7
8b	ASSR	Signal + Noise	240	65	240	0.0	0.0	0.0	100.0	0.0
8b	ASSR	Noise	240	65	14	94.2	97.5	3.3	96.7	44.0
9a	ASSR	Signal + Noise	276	85	260	5.8	12.0	6.2	93.5	5.1
9a	ASSR	Noise	276	85	21	92.4	97.5	5.1	94.9	47.2
10	ASSR	Signal + Noise	750	260	323	56.9	60.5	3.6	95.9	25.4
10	ASSR	Noise	750	260	40	94.7	93.1	-1.6	97.9	39.8
8c	ASSR	Signal + Noise	480	65	458	4.6	6.2	1.6	98.3	2.3
8c	ASSR	Noise	480	65	26	94.6	97.3	2.7	97.3	45.6
9b	ASSR	Signal + Noise	184	85	148	19.6	27.2	7.6	91.8	12.7
9b	ASSR	Noise	184	85	13	92.9	97.3	4.4	95.7	46.8



Sup. Figure 1. Robust performance of the Fmpi under stationary and non-stationary noise in ABR simulations. a Examples of ABR simulations under stationary (left) and non-stationary (right) background noise. The template ABR (Target) is shown in red. Weighted average waveforms containing both the ABR template and noise are depicted in light blue (Signal + Noise), whilst averages obtained without the target ABR (Noise) are shown in green, with the standard error of the mean indicated by the shaded green area. **b** Nominal versus observed sensitivity for clinically relevant thresholds (95-100%), generated by pooling noise-only data across all epochs. The diagonal solid black line represents the ideal unbiased expectation. Insets show the average bias across nominal sensitivities; a positive bias indicates higher observed sensitivity than nominal, whilst a negative bias indicates the opposite. **c** False response detection rates from the noise-only data at a nominal sensitivity of 95%, pooled across all epochs. **d** Response detection rates (Signal + Noise) as a function of the number of epochs. Each epoch includes 1000 simulations. **e** AAUC for the clinically relevant nominal sensitivities. Shaded areas represent 95% confidence intervals. Error bar lengths have been adjusted for visual post hoc comparisons (Holm FWER corrected); non-overlapping segments indicate significant differences, whilst overlapping segments indicate non-significant differences. The horizontal solid black line indicates the chance level of the AAUC. The different methods are indicated by colour across all panels.



Sup. Figure 2. Robust performance of the Fmpi under stationary and non-stationary noise in ASSR simulations. a Examples of 41 Hz ASSR simulations under stationary (left) and non-stationary (right) background noise. The template ASSR (Target) is shown in red. Weighted average waveforms containing both the ASSR template and noise are depicted in light blue (Signal + Noise), whilst averages obtained without the target ASSR (Noise) are shown in green, with the standard error of the mean indicated by the shaded green area. All detectors employed 3 harmonics (six features) for statistical assessments. **b** Nominal versus observed sensitivity for clinically relevant thresholds (95-100%), generated by pooling noise-only data across all epochs. The diagonal solid black line represents the ideal unbiased expectation. Insets show the average bias across nominal sensitivities; a positive bias indicates higher observed sensitivity than nominal, whilst a negative bias indicates the opposite. **c** False response detection rates from the noise-only data at a nominal sensitivity of 95%, pooled across all epochs. **d** Response detection rates (Signal + Noise) as a function of the number of epochs. Each epoch includes 1000 simulations. **e** AAUC for the clinically relevant nominal sensitivities. Shaded areas represent 95% confidence intervals. Error bar lengths have been adjusted for visual post hoc comparisons (Holm FWER corrected); non-overlapping segments indicate significant differences, whilst overlapping segments indicate non-significant differences. The horizontal solid black line indicates the chance level of the AAUC. The different methods are indicated by colour across all panels.



Sup. Figure 3. Analytical DOF and critical F-values for various stationary noise types and bandwidths. a DOF v_y and 95% critical F-value for a fixed $v_{RN} = 256$ for cortical-like analysis bandwidths, using an analysis window of 500 ms. The low-pass (LP) and high-pass (HP) ideal cutoff frequencies are indicated on the y- and x-axes, respectively. Each panel corresponds to a noise with a PSD of $1/f^p$ with p ranging from 0 (white noise) to 1 (pink noise). The range of the corresponding values is shown by the colour bar. **b** Same as in panel **a**, but for ABR-like analysis bandwidths, using an analysis window of 10 ms. The green markers for white (p=0) and pink (p=1) noise in panel **b** correspond to analytical DOF values of $v_y = 160.0$ (white noise) and $v_y = 13.0$ (pink noise), respectively. The empirical values reported by Elberling and Don[24] for the same bandwidth and noise types were 160 and 15, respectively.