**Controlling test specificity for auditory evoked response detection using a frequency domain bootstrap**

**Chesnaye M.A1., Bell S.L2., Harte J.M3., & Simpson D.M4**

1Institute of Sound and Vibration Research, Faculty of Engineering and the Environment, University of Southampton, [UK. mac1r19@soton.ac.uk](mailto:UK.mac1r19@soton.ac.uk)

2Institute of Sound and Vibration Research, Faculty of Engineering and the Environment, University of Southampton, [UK. s.l.bell@soton.ac.uk](mailto:UK.mac1r19@soton.ac.uk)

3Eriksholm Research Centre, Snekkersten, Denmark. jmha@eriksholm.com

4Institute of Sound and Vibration Research, Faculty of Engineering and the Environment, University of Southampton, UK. ds@isvr.soton.ac.uk

**Highlights**

1. The frequency domain bootstrap (FDB) is introduced to ABR detection
2. FDB modifications are proposed to account for heteroskedasticity and non-smooth PSDs
3. The FDB and modified FDB approaches outperform parametric response detection methods

**Abstract**

*Background*: Statistical detection methods are routinely used to automate auditory evoked response (AER) detection and assist clinicians with AER measurements. However, many of these methods are built around statistical assumptions that can be violated for AER data, potentially resulting in reduced or unpredictable test performances. This study explores a frequency domain bootstrap (FDB) and some FDB modifications to preserve test performance in serially correlated non-stationary data.

*Method*: The FDB aims to generate many surrogate recordings, all with similar serial correlation as the original recording being analysed. Analysing the surrogates with the detection method then gives a distribution of values that can be used for inference. A potential limitation of the conventional FDB is the assumption of stationary data with a smooth power spectral density (PSD) function, which is addressed through two modifications.

*Comparisons with existing methods:* The FDB was compared to a conventional parametric approach and two modified FDB approaches that aim to account for heteroskedasticity and non-smooth PSD functions. Hotelling’s (HT2) test applied to auditory brainstem responses was the test case.

*Results*: When using conventional HT2, false-positive rates deviated significantly from the nominal alpha-levels due to serial correlation. The false-positive rates of the modified FDB were consistently closer to the nominal alpha-levels, especially when data was strongly heteroskedastic or the underlying PSD function was not smooth due to e.g. power lines noise.

Conclusion: The FDB and its modifications provide accurate, recording-dependent approximations of null distributions, and an improved control of false-positive rates relative to parametric inference for auditory brainstem response detection.

**1. Introduction**

Auditory evoked responses (AERs) are small electrical potentials generated along the auditory pathway following an auditory stimulus (Picton, 2011). They are typically recorded non-invasively from the scalp using the electroencephalogram (EEG), and are important diagnostic tools in the clinic. Some of their main applications include (i) hearing threshold estimation, hearing screening and hearing aid fitting verification ([Punch](https://www.ncbi.nlm.nih.gov/pubmed/?term=Punch%20S%5BAuthor%5D&cauthor=true&cauthor_uid=27587921) et al, 2016; Sininger et al, 2018), (ii) monitoring changes in neurological activity (e.g. intra-operative monitoring; [Møller](https://www.ncbi.nlm.nih.gov/pubmed/?term=M%C3%B8ller%20AR%5BAuthor%5D&cauthor=true&cauthor_uid=6435065), 1988), and (iii) diagnosing or confirming various neurological disorders (Duncan et al, 2009).

One of the main challenges for AER-related applications is the low signal-to-noise ratio (SNR) of the AER, which makes discriminating the AER from the EEG background activity a challenging task. To improve the SNR, many stimuli are presented to the subject, and the intervals following stimulus onset (henceforth named “epochs”) are averaged to reduce noise. Response detection is then usually achieved by highly trained individuals, who visually inspect the averaged waveforms. The examiners can also be assisted by objective detection methods (e.g. Elberling & Don, 1984; Cebulla et al, 2006; Chesnaye et al, 2018), which have the goal to confirm or refute the presence of an evoked response. Ultimately, the aim for the detection method is to reduce subjectivity, and to provide an overall more reliable, sensitive, and efficient (i.e. shorter in duration) test.

Some of the more well-known methods for assisting examiners with response detection are the Fsp (Elberling & Don, 1984), the magnitude squared coherence (Dobie & Wilson, 1989), the spectral F-test (Valdes, 1997), the Hotelling’s T2 test (Golding et al, 2009) and correlation coefficients (e.g. Valderrama et al, 2014). These methods are all parametric, i.e. they use theoretical distributions to evaluate test significance, and thus make various assumptions regarding the EEG data. When these assumptions are violated, the assumed null distributions can be inaccurate and decreased test sensitivities and/or unexpected test specificities can result.

As an alternative to the parametric approach, some authors have also suggested resampling methods to evaluate test significance, e.g. the moving block bootstrap (MBB; Lv et al, 2007) and the permutation test (Maris & Oostenveld, 2007) have been proposed. These methods strive to approximate the null distributions of the test statistic, as opposed to assuming them, achieved by constructing many “surrogate” or “pseudo” recordings by randomly resampling continuous blocks of EEG data from within the original recording. The surrogate recordings are then analysed with the detection method to generate a distribution of test values, which is assumed to approximate the null distribution of the test statistic, representing “no AER present”.

Resampling methods are appealing, firstly because mathematical tractability of the underlying distributions is no longer required, which gives the user a large amount of freedom when choosing which test statistic to use for response detection. The resampled distributions can potentially also be more accurate than the theoretical sampling distributions, albeit under the condition that the surrogates emulate important data characteristics that impact on the test statistic. Current resampling methods for AER detection (Lv et al, 2007; Maris & Oostenveld, 2007) emulate the recording mean and variance along with serial correlation within epochs, but disrupt serial correlation between epochs. The latter may result in a mismatch between the original and the surrogate recordings, and a biased approximation of the null distribution underlying the test statistic.

The issue of serial correlation within epochs is well-known for AER detection, and was the incentive for the conservative design of the Fsp statistic (Elberling & Don, 1984). The impact of serial correlation between epochs, however, has received less attention. Geisler (1960) visually inspected the autocorrelogram of 8-600 Hz band-pass filtered EEG and observed autocorrelations up to at least 50 ms, whereas others have detected significant autocorrelations for much longer periods of time, up to 3 seconds or more depending on the spectral content of the data (Neely & Pepe, 1997; Victor & Mast, 1991; Mast & Victor, 1991). Serial correlation between epochs has also been shown to inflate FPRs in a sequential HT2 test (Chesnaye et al 2019), and is hence a potential concern for AER detection. The first aim for this work was therefore to provide a more rigorous assessment of serial correlation, and to quantify its impact in terms of increased or decreased false-positive rates (FPRs) as a function of commonly used test parameters. The assessment was carried out for auditory brainstem response (ABR) data using the parametric Hotelling’s T2 test as detection method, which has previously shown high test sensitivity for ABR detection (Vanheusden et al, 2018; Chesnaye et al, 2018).

A second aim for this work was to obtain good control over the FPR, relative to the conventional parametric approach, by adopting a suitable bootstrap that emulates serial correlation within the ABR data. There is a rich literature on randomisation and resampling methods for correlated time series (for references, see Lahiri, 2003; Kreiss & Paparoditis, 2011; see also the discussion). This work opted to use the frequency domain bootstrap (FDB) approach, which efficiently generates stationary surrogates with a common power spectral density (PSD) function and is readily amenable to adaptation according to signal-specific requirements. Perhaps the main appeal of the FDB is that bootstrap resampling is applied in the frequency domain, which allows serial correlation to be emulated without having to estimate the time domain correlation structure (Paparoditis 2002).

A potential drawback of the FDB is that data is assumed to be stationary with a smooth PSD function. This is a concern for ABR data as some noise sources can be intermittent (potentially introducing non-stationarities to the data), and/or can introduce sharp peaks to the PSD function, e.g. due to power lines interference. In addition to the conventional FDB, this work therefore also explores two FDB modifications that aim to (1) reduce the impact of non-stationarities by accounting for non-stationary variance (hetereoskedasticity), and (2) obtain an improved estimate of the PSD function by effectively dealing with sharp peaks in the PSD (sections 2.2.2 and 2.2.3, respectively). Other non-stationary data characteristics, and some alternative methods for generating stationary and non-stationary surrogates, are briefly considered in the discussion.

To summarise, the aims for this work were (1) to quantify the impact of serial correlation in terms of increased or decreased FPRs when using the parametric Hotelling’s T2 (HT2) test for ABR detection, and (2) to obtain an improved control over the FPR by adopting the FDB and two modified FDB approaches. Data for the assessment consisted of a large amount of no-stimulus EEG recordings from 16 adults, an ABR threshold series recorded from 12 normal-hearing adults, and simulations. Various parameter choices underlying the performance of the FDB and FDB modifications, along with some limitations and potential applications, are considered in the discussion.

**2. Methods**

This section describes the data (Section 2.1), the FDB and its modifications (Section 2.2), the HT2 test (Section 2.3), and the procedures for evaluating test specificity (Section 2.4) and test sensitivity (Section 2.5).

**2.1. Data**

***EEG background activity***

Recordings of EEG background activity (no stimulus was used) were previously collected from 12 male and 4 female subjects (Madsen et al, 2018) under various test conditions. In this study, data from test conditions “still” and “sleep” were included in the analysis, where subjects were asked to lie still with there eyes closed, but not to fall asleep, or to lie still with their eyes closed and try to fall asleep, respectively. Measurements were obtained using a Compumedics Neuroscan II EEG amplifier at a sampling rate of 20 kHz (subsequently anti-alias filtered and downsampled offline to 5 kHz) with silver–silver chloride (Ag/AgCl) electrodes placed on the left mastoid (active), the right cheek (ground) and the upper forehead (reference). The electrode impedances remained below 1 kΩ throughout the recording for all subjects. There were a total of 96 recordings available for the analysis with a mean recording time of ~4.2 minutes (sd. 1.54 minutes) and a total recording time of ~7 hours prior to artefact rejection. Data were band-pass filtered offline from *fc* to 1500 Hz using 3rd-order Butterworth filters, where *fc* was the high-pass cut-off frequency, which was later varied from 30 to 100 Hz in steps of 5 Hz. The high-pass cut-off frequency *fc* was varied as it impacts on serial correlation within the EEG. The range chosen covers values commonly used in the clinic, and provides a means to test the efficacy of the FDB under different sample correlation structures (see Section 2.4 for more details). Finally, artefact rejection was applied offline by discarding all epochs that contained absolute maximum values larger than 10 µV, as recommended by the British Society of Audiology (BSA; Lightfoot et al, 2019).

***ABR threshold series***

ABR data were previously recorded from 6 female and 6 male subjects (aged 18 to 30 years) with normal hearing (< 20 dB HL for 250, 500, 1000, 4000, and 8000 Hz tones), and is described in detail in Lv et al (2007). Measurements were obtained at a sampling rate of 10 kHz (anti-alias filtered and down-sampled offline to 5 kHz) using a Cambridge Electronic Design (CED, Cambridge, UK) micro 1401 data acquisition unit along with a CED 1902 amplifier. ABRs were recorded with the active electrode placed at the vertex, a reference electrode at the nape of the neck and a ground electrode placed at mid-forehead. Electrode impedances remained below 5 kΩ throughout the recording. The stimulus for evoking the ABR was a 100 µs click, delivered at a rate of 33.33 Hz at 50, 40, 30, 20, 10, and 0 dB sensation levels (SL), i.e. relative to the behavioural hearing levels of each participant. Two recordings were made per dB SL condition, with a minimum of 2000 stimuli being presented per recording (stimulus presentation was just over one minute per recording). Data were processed in a manner similar to the background EEG background activity described above, i.e. band-pass filtered offline from *fc* to 1500 Hz using 3rd-order Butterworth filters, where *fc* again ranged from 30 to 100 Hz, in steps of 5 Hz. Finally, artefact rejection was applied by discarding all epochs (the ~30 ms segments following stimulus onset) that contained maximum absolute values larger than 10 µV. The median ensemble size after artefact rejection was 1458 epochs, with first and third quantiles equal to 810 and 1808 epochs, respectively. Ethics approval was obtained from Southampton University’s Ethics and Research Governance Committee.

**2.2. The frequency domain bootstrap**

Consider EEG recording  (*n*=1, 2, …, *N*), where *N* is the total number of samples in the recording. This is analysed with some detection method, giving the test statistic T, e.g. T might be the Fsp or the statistic (see also Section 2.3). The goal for the FDB, as used in the current work, is to estimate the corresponding p value, or a probability that T arose under the null hypothesis of “no ABR present”. To do so, the null distribution underlying T is required, which is approximated by the FDB by generating many surrogate recordings with no response present (further described below) and recalculating T from each of these surrogates. It is worth emphasising here that the surrogates are all derived from the specific recording being analysed, i.e. each recording is associated with its own unique bootstrapped distribution, which allows recording-dependent signal (and noise) characteristics (including serial correlation) to be taken into account. The alternative approach of first collecting a large amount of no-stimulus data from a cohort of participants and then using this data to approximate the null distribution would need to assume that data characteristics (e.g. non-stationarity and serial correlation) in the a priori recorded no-stimulus data is mirrored in each of the recordings being analysed, which would be a risky assumption for EEG data.

In what follows, a description is provided for the conventional FDB (Kreiss & Paparoditis, 2011). Detailed descriptions can also be found in Dahlhaus & Janas (1996) and Paparoditis (2002). Two FDB modifications are then also presented, after which implementation details and an illustrative example are provided.

**2.2.1 The conventional frequency domain bootstrap**

The two main components of the FDB include the sample periodogram and the true power spectral density (PSD) function of the time series . The latter needs to be estimated from  (see below) and plays a key role in the performance of the FDB. Starting with the sample periodogram , this is given by (Paparoditis, 2002):

 Eq. 1

Where *N* is the length of the recording  and  is the frequency. The sample periodogram is typically evaluated at frequencies Hz (*j*=1, 2, …, ) where *fs* is the sampling rate. The core idea underlying the FDB is that  are asymptotically independent exponentially distributed random variables for *j* = 2, …, - 1 (Dahlhaus & Janas, 1996; see also theorem 10.3.2 in Brockwell & Davis, 1991), which allows bootstrap resampling (Efron & Tibshirani, 1993) to be applied to the rescaled sample periodogram ordinates. In particular, the following bootstrap procedure has been proposed (Dahlhaus & Janas, 1996):

1. Define the “Studentized periodogram ordinates” as  where is an estimate of .
2. Define the rescaled periodogram ordinates as 
3. Generate random surrogate PSDs using  where is randomly resampled (with replacement) from the empirical distribution . The aim of this step is to ensure that the surrogate periodograms have the expected random variability.

The surrogate periodograms can then be transformed into time series using a “time frequency toggle” approach (Kirch & Politis, 2011):

1. Transform the surrogate periodograms into magnitudes using  where  is the spectral resolution and assign a random phase (uniformly sampled from the [0, 2π] interval) to each frequency.
2. Take the inverse FFT to obtain time domain surrogates . Note that, in order to obtain a real signal, the FFT coefficients should be forced to provide Hermite symmetry around the Nyquist frequency.

The procedure can be further simplified by exploiting priori knowledge of the distribution of , which is assumed to follow a standard exponential distribution with a mean of one (Dahlhaus & Janas, 1996). To clarify, note that the power at frequency  can be expressed as a sum of the squared real and imaginary parts of the FFT, which, being a sum of squares, follows a scaled  distribution with 2 degrees of freedom. The  distribution is then related to the exponential distribution through  where *Exp*(1) is an exponential distribution with a mean of one (Leemis & Mcquestion, 2008). In the current work, were sampled from *Exp*(1), as opposed to the empirical distribution ..

As mentioned previously,  needs to be estimated from  and plays a crucial role in the performance of the FDB. The PSD function can be estimated robustly using the well-known Welch method (Welch, 1967) with overlapping Hanning windows to provide spectral smoothing. To ensure that has the correct amplitude, it is also rescaled, such that its mean power, given by , is equal to the mean square of .

**2.2.2 Modification I. FDB Envelope - accounting for heteroskedasticity**

The conventional FDB assumes that  has a single underlying PSD function, and hence that data is stationary throughout the recording. To reduce the impact of non-stationarities on the subsequent ABR detection,  is rescaled by its root-mean-square (RMS) envelope to make it more stationary prior to PSD estimation, i.e.  (using sample by sample division) whereis the estimated RMS envelope and  is the rescaled signal. The FDB is then applied to the more stationary  recording. The surrogates  are later multiplied by  to reintroduce heteroskedasticity. Further implementation details on how  was estimated are provided in section 2.2.4 below, whereas other types of non-stationarities and some alternative methods are briefly considered in the discussion.

**2.2.3 Modification II. FDB Peaks - accounting for non-smooth PSD functions**

An additional limitation for the FDB is that is generally ill-suited to capturing sharp peaks in , such as those due to power lines noise and its harmonics. This is due to the Hanning window in the Welch method (Welch, 1967), which introduces a smoothing effect on , i.e. peaks in  will tend to be attenuated in . In addition, the attenuated power will be smeared across the neighbouring frequencies, resulting in an over-estimated power for the frequencies surrounding the peaks.

Modification II aims to circumvent peaks in when estimating the PSD function, achieved by first interpolating across the peaks in , and then using the inverse FFT to go back to the time domain, giving modified recording . A modified PSD function, say , is then estimated from . A potential risk with this approach is that the interpolation can be inaccurate due to random variation in . A second step is therefore included: Going back to the original sample periodogram , the original peaks in  are now replaced with the noise floor estimated in . The modified  is then transformed to the time domain, giving a second modified recording , after which the PSD function is again estimated, now from . The original peaks in  are later substituted back into the surrogate periodograms.

**2.2.4 An illustrative example and implementation details**

In this example, the FBD and its modifications were applied to a short segment (4.43 seconds) of clearly non-stationary EEG (Fig.1a). As shown in Fig1.b, this data contained an artefact around the two second mark, which is shown in gray along with a ±10 µV artefact rejection threshold (shown as red dotted lines). It is important to consider artefact rejection as it introduces discontinuities in , which need to be handled appropriately.

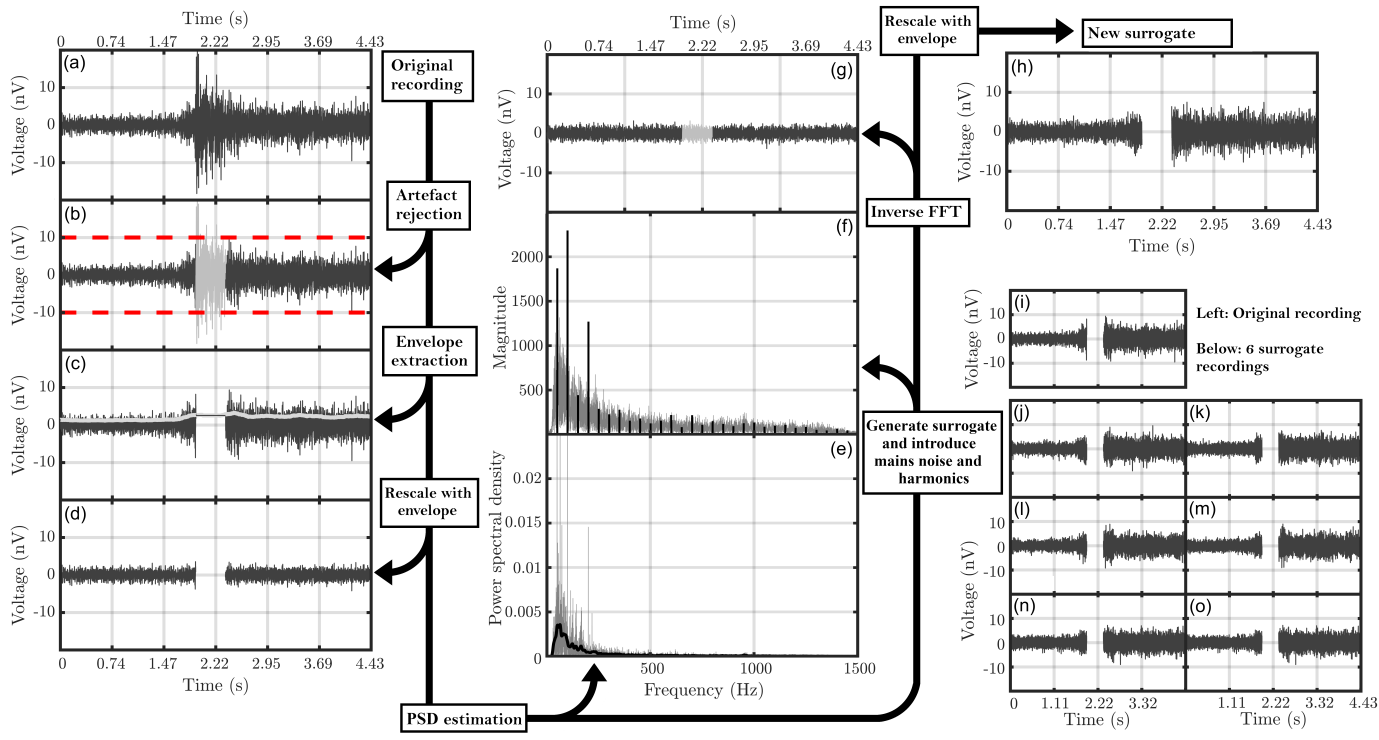
When using the FDB modification “FDB Envelope” (Section 2.2.2), the first step is to estimate , achieved by sliding a rectangular 200 ms window across  and calculating the RMS of the windowed EEG for all *n*. The resulting envelope is shown in Fig.1c as a thick gray line, whereas the rescaled (more stationary) time series  is shown in Fig.1d. To prevent artefacts from distorting , the rejected interval of  (the gray line in Fig.1b) was replaced with the RMS of its two neighbouring intervals (100 ms on both sides) prior to extracting . The envelope  was also low-pass filtered at 5 Hz using a 3rd order Butterworth filter.

In the next step, was estimated by shifting a 200 ms Hanning window with 80% overlap across the recording, and then averaging across the windowed PSDs. Note that PSD function estimation will differ, depending on whether the conventional FDB or one of the modified FDB approaches is used. When using the “FDB Peaks” modification, it was assumed that the underlying PSD contained sharp peaks due to power lines noise at 50±0.5 Hz and all harmonics, up to 1500±0.5 Hz. Note that the ±0.5 Hz intervals were necessary as the power lines were not always located exactly at 50 Hz due to small variations in mains frequency and potentially small inaccuracies in analogue-to-digital converter frequencies. The estimated PSD function for this example is shown in Fig1.e as a thick black line, and is a smoothed version of , shown in gray.

In the current work, care was taken to avoid discontinuities due to artefact rejection by shifting the 200 ms Hanning window across just the continuous segments of EEG. For this example, all continuous segments of EEG were longer than 200 ms. However, in some recordings, artefact rejection occasionally resulted in EEG segments with durations shorter than 200 ms, i.e. too short to extract a single 200 ms window of continuous EEG. These segments were not used for PSD estimation, although additional analysis (details not presented) suggests that these shorter segments can be concatenated with their longer neighbours with little adverse effect.

Random surrogate periodograms were then generated using step (3) in Section 2.2.1. When using “FDB Peaks”, a priori identified peaks in were substituted back into the surrogate periodograms (Fig.1f), after which time domain surrogates were generated by transforming the surrogate perdiodograms to the time domain using the inverse FFT, albeit after transforming the spectral densities to magnitudes and assigning random phases . An example of a surrogate time series is shown in Fig.1g. Note also that the surrogates should contain the same discontinuities due to artefact rejection as , else serial correlation between  and the surrogates will still differ. The intervals discarded in  should therefore also be discarded in the surrogates (Fig1.g), even though these contain no artefacts.

Finally, when using the FDB modification “FDB Envelope”, the stationary surrogates were rescaled with the envelope, i.e.  (sample by sample multiplication). An example of a final surrogate time series is shown in Fig.1h, with six more being presented in Fig.1j to Fig.1o.



**Figure 1.** The FDB and modified FDB procedures for generating EEG surrogate data. First, for some EEG recording (panel a), standard pre-processing and artefact rejection was applied (panel b). When using the FDB modification “FDB Envelope”, the RMS envelope was extracted (panel c), which was used to rescale the EEG record (panel d). Next, the PSD function was estimated (panel e), and used as a model for generating many additional surrogate PSDs. Note that the PSD-estimation procedure will differ, depending on whether the original FDB or one of the modified FDB approaches were used. The random surrogate PSDs were then converted to magnitudes (panel f), and the 50 Hz mains noise and harmonics added back into the surrogates (when using the “FDB Peaks” modification), after which the surrogates were transformed to the time domain (panel g). The intervals discarded in  due to artefact rejection were then also discarded in the surrogates. Finally, when using the “FDB Envelope” modification, the surrogates were rescaled with the previously estimated RMS envelope to re-introduce non-stationarities in the sample variance (panel h). For demonstration purposes, 6 additional surrogates are shown in panels j to o.

**2.3. The one-sample Hotelling’s T2 test**

The one-sample Hotelling’s T2 test (HT2; Hotelling, 1931) is the multivariate equivalent to Student’s one-sample t-test, and can be used to test whether Q feature means are significantly different from Q hypothesized values. When used for AER detection, Q features (defined below) are extracted from each epoch, giving an MxQ dimensional matrix of features, say **V**,where M is the number of epochs. The Q feature means are then found by taking the means down the Q columns of **V**. The statistic itself is a weighted sum of the Q feature means where the weights are determined by the variances and covariances of the features (Rencher, 2001):

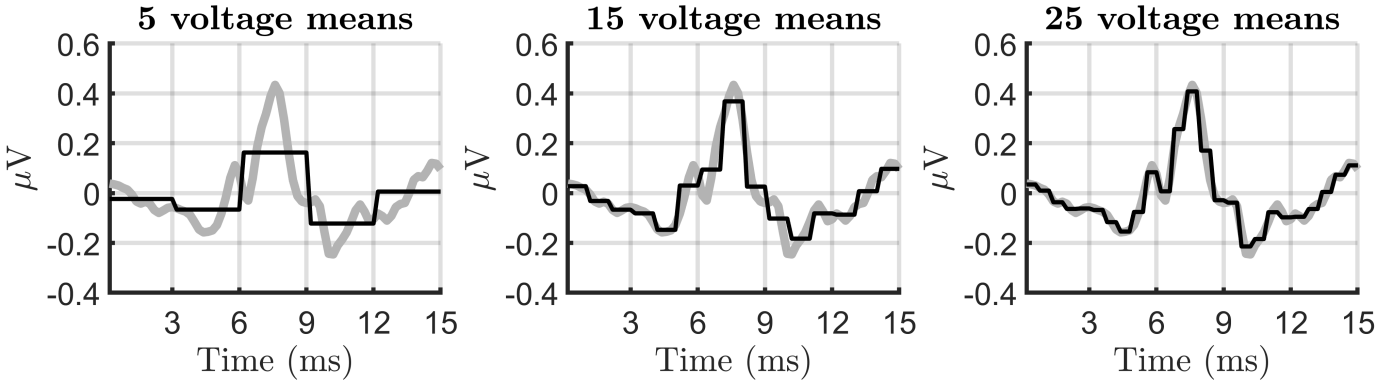
 Eq. 2

where  is the Q-dimensional vector of feature means,  is the Q-dimensional vector of hypothesized values to test against (all zeros under the null hypothesis of no AER present),  is the inverse of the covariance matrix of **V**, and ** superscript denotes Hermitian transpose. The statistic can then be transformed into an F-statistic using, which is F-distributed with Q and M-Q degrees of freedom under H0 (Rencher, 2001), albeit under the assumption that **V** is stationary, Gaussian-distributed, and independent across rows (i.e. independent epochs). This distribution will be referred to as the theoretical sampling distribution (or simply F-distributions), which may be contrasted to that obtained from the FDB. In subsequent sections, the significance of  is evaluated using either F-distributions or bootstrapped distributions generated by the FDB and its modifications.

*Features: Voltage-means*

In this work, features consist of mean voltages, taken across short time intervals within epochs (Golding et al, 2009), henceforth denoted as ‘voltage-means’. When a response is present, the Q feature means ( in Eq. 1) will tend to be non-zero, whereas when a response is absent, their expected values are zero as the mean of the recording is shifted to baseline by the high-pass filter during the pre-processing stage. The Q-hypothesized values under H0 to test against ( in Eq. 1) are therefore all zero.

With respect to the number of voltage means Q to extract from each epoch, this will depend on the expected ABR waveform along with serial correlation within the data. In particular, when Q is small, consecutive peaks and troughs in the ABR waveform may cancel out, resulting in a loss of information (see Figure 2 below), whereas when Q is large, consecutive voltage means can be highly correlated, resulting in an increasingly (i.e. increasing with both Q and serial correlation) ill-conditioned matrix **S**. For this study, a total of Q=25 voltage means was used, as this has previously shown good performance in ABR detection (Chesnaye et al, 2018).



**Figure 2.** A demonstration of feature extraction and the potential loss of information when the number of voltage means Q is too small. The thick gray lines show a coherently averaged waveform that contained a clear ABR, whereas the thinner black lines show the voltage means, superimposed on the EEG segments from which they were obtained. For Q = 5 and Q = 15, the EEG segments to average across are relatively wide, resulting in a loss of information. Note that this example applied feature extraction to the coherently averaged epoch for demonstration purposes only. In practice, feature extraction is applied to the individual epochs.

**2.4. Specificity**

The goal for the specificity assessment was to evaluate the FPRs when conducting statistical inference on the statistic using either the theoretical sampling distribution (F-distributions), the FDB, or the two FDB modifications. Data for the assessment consisted of the no-stimulus EEG recordings and simulations.

**2.4.1 Simulations**

Data for the simulations consisted of stationary colored noise, generated by filtering Gaussian White Noise with all-pole filters where the poles of the filters were given by the parameters of 60th-order autoregressive (AR) models. The AR models were estimated from the no-stimulus EEG recordings using the Modified Covariance method (Marple, 1987), with a new AR model being fit to each recording. All simulated recordings were band-pass filtered from 30-1500 Hz using 3rd-order Butterworth filters, and re-structered into ensembles of 189 epochs using a hypothetical stimulus rate of 47.17 Hz, giving recordings of ~4 seconds.

*Simulations I: heteroskedasticity*

To simulate different degrees of non-stationary variance, or heteroskedasticity, half of the epochs within each ensemble (epochs 95-189) were rescaled so that a certain standard deviation (s.d) was obtained, which was varied from 2 to 10 uV, in steps of 0.2 uV. The s.d of the first half of the ensemble (epochs 1-94) was always fixed at 2 uV. A total of 10,000 recordings were simulated, per test condition.

*Simulations II: non-smooth PSD functions*

To simulate non-smooth PSD functions (in this case due to power lines noise), a continuous 50 Hz sinusoid was added to each simulated recording. The amplitude of the sinusoid was varied from 0 to 5 uV, in steps of 0.25 uV. A total of 10,000 recordings were simulated, per test condition.

**2.4.2. Subject data: no-stimulus recordings**

The subject-recorded no-stimulus data were pre-processed using a range of high-pass cut-off frequencies *fc* and (hypothetical) stimulus rates, as described in section 2.1. These parameters were varied so that test performance could be evaluated for different sample correlation structures. In particular, note that serial correlation is determined primarily by the dominant frequency (the frequency with the largest amplitude) in the data, which will tend to be the lowest frequency due to the approximate  spectrum (with; Pritchard, 1992) of the EEG background activity, and is thus determined primarily by the high-pass cut-off frequency *fc*. The distance between epochs then determines the extent to which consecutive analysis windows (typically the 0-15 ms windows within epochs) are correlated. Together, the high-pass cut-off frequency and stimulus rate have a large impact on the correlation structure underlying the data to be analysed.

The pre-processed no-stimulus recordings were structured into ensembles of M epochs, after which the initial 15 ms windows of the ensembles were used to calculate the statistic. The ensemble size M was furthermore always chosen such that the total amount of artefact free data per ensemble was approximately 4 seconds, e.g. when using a stimulus rate of 28.6 Hz (approximately 35 ms epochs) the ensemble contained 115 artefact free epochs, whereas when using a stimulus rate of 66.67 Hz (approximately 15 ms epochs), this was 267.

**2.4.3. Post hoc analysis: confidence intervals using the binomial distribution**

To test whether the FPRs deviated significantly from an α-level of 0.05 (where α is the significance level of the test), two-sided 99% confidence intervals were constructed using Binomial distributions. In particular, a Bernoulli distribution was constructed from B Bernoulli trials where B was the number of tests performed, and where the probability of a single successful Bernoulli trial was set to α=0.05, i.e. the theoretical probability of a false-positive. For the subject data, the number of tests B ranged from ~3000 to ~4500, depending on the high-pass cut-off frequency which impacted on the amount of artefact-free data. The corresponding 99% confidence intervals ranged from ~[0.0402, 0.0607] to ~[0.0419, 0.0587] for B=3000 and B=4500, respectively. For the simulations, B was set to 10,000, giving 99% confidence intervals of [0.0445, 0.0557] for the expected FPR of 0.05.

**2.5. Sensitivity**

For the sensitivity assessment, the aim was to evaluate the detection rates when using the  statistic where test significance was evaluated using either the theoretical sampling distribution (F-distributions), the FDB, or the two modified FDB approaches. Data for the assessment were the subject-recorded ABR threshold series (see Section 2.1) and simulations.

**2.5.1. Simulations**

Data for the simulations consisted of the previously described stationary coloured noise (Section 2.4.1) for emulating the EEG background activity, along with rescaled ABR templates for simulating a response. The ABR templates were given by the coherently averaged ABR waveforms from the subject-recorded ABR threshold series, under the condition that the coherent average contained a clear response. The latter was determined by an experienced audiologist, who used BSA guidelines (Lightfoot et al, 2019). When simulating an ABR, one of the templates (there were 27 in total) was selected at random, rescaled, and added to all epochs within the ensemble in question. The scaling factor was furthermore chosen to obtain a certain SNR, which was varied from -40 dB to -16 dB. A total of 5,000 recordings were thus simulated, per SNR. The initial 0-15 ms windows of the ensembles were then used to generated a statistic, and inference was carried out using F-distributions, the FDB, or the two FDB modifications.

**2.5.2. ABR threshold series**

For the subject-recorded ABR threshold series, the initial 1-15 ms post-stimulus windows of the ensembles were used to generate a statistic. Note that the first ms was excluded to avoid potential contaminations from a stimulus artefact. Inference was again carried out using F-distributions, the FDB, and the FDB modifications.

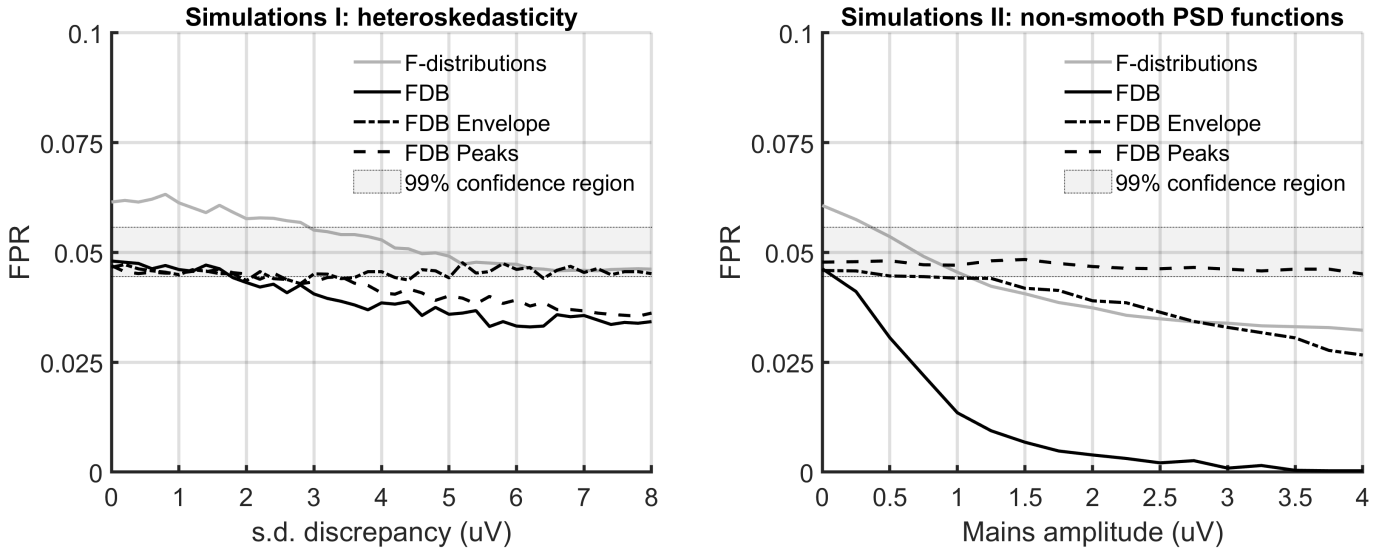
**3. Results**

This section presents the results from the specificity assessment (Section 3.1) and the sensitivity assessment (Section 3.2).

**3.1. Specificity**

**Simulations:** Results from Simulations I and II are presented in Figure 3, and demonstrate how heteroskedasticity (Simulations I, left panel) and non-smooth PSD functions (Simulations II, right panel) can lead to significantly inflated or reduced FPRs. Results from Simulations I show that test performance became increasingly conservative (FPRs < α) as non-stationarity in the variance became more severe with all methods, except the FDB Envelope approach. Results from Simulations II show that test performance also became increasingly conservative for all methods, except the FDB Peaks approach. Note that a lower than expected FPR is not necessarily desirable, as it goes hand in hand with a reduced test sensitivity. Note also that in practice, violations to multiple assumptions may occur simultaneously, potentially resulting in additional inflations to error rates. Alternatively, multiple violations may end up cancelling each other out, i.e. two wrongs may end up making a right.

It is worth emphasising that the simulated 50 Hz power lines noise in Simulations II impacted on the FPRs when using a (hypothetical) stimulus rate of 47.17 Hz. This can be attributed to the smoothing effect from the sliding Hanning window in Welch’s method (see also section 2.2.3), which smeared the power of the peaks across their neighbouring frequencies, resulting in an over-estimated power for these frequencies, and thus a conservative test performance. Note that a conservative test performance was also observed when using F-distributions, which can likely be attributed to additional independence violations, i.e. as the amplitude of the simulated power lines noise was increased, it became the largest frequency in the data, and increasingly dominated the underlying correlation structure.

****

**Figure 3.** False-positive rates (FPRs) from the specificity assessment as a function of (1) the difference in the standard deviations calculated from the first and second half of each ensemble (i.e. from epochs 1-94 and 95-189, respectively), which aimed to emulate various degrees of non-stationary variance (heteroskedasticity), and (2) the simulated magnitude of a 50 Hz sinusoid (Simulations II, right panel), here emulating a 50 Hz power lines noise. Data were analysed with the T2 statistic and test significance was evaluated using either F-distributions, the original frequency domain bootstrap (FDB) approach, or either the “FDB Envelope” and “FDB Peaks” modifications. The 99% confidence intervals for the expected FPR of α=0.05 are also shown. These simulations used a hypothetical stimulus rate of 47.17 Hz.

**Subject-recorded no-stimulus data:** The FPRs (α=0.05) from the specificity assessment are plotted as a function of the high-pass cut-off frequency *fc* and the (hypothetical) stimulus rate in Figure 4 for F-distributions (panel A), the original FDB (panel B), and the two modified FDB approaches (panels C and D). FPRs that deviate significantly (p<0.01) from α=0.05 are shown in red (FPR > α) and blue (FPR < α) whereas FPRs that fall within the expected 99% confidence intervals are shown in green. The following observations are made:

1. Independence violations

When inspecting panel (A) of Figure 4, note the “wave-like” pattern in FPRs, starting from the upper left corner, moving towards the lower right. These deviations are a function of the high pass cut-off frequency and stimulus rate, which strongly suggests that they are due to serial correlation in the data, which violated the independence assumption between epochs underlying the parametric HT2 test (although stationarity and Gaussianity violations likely also play a role). Note also that these deviations are absent in panels B-D, as serial correlation has been accounted for by the FDB and modified FDB approaches.

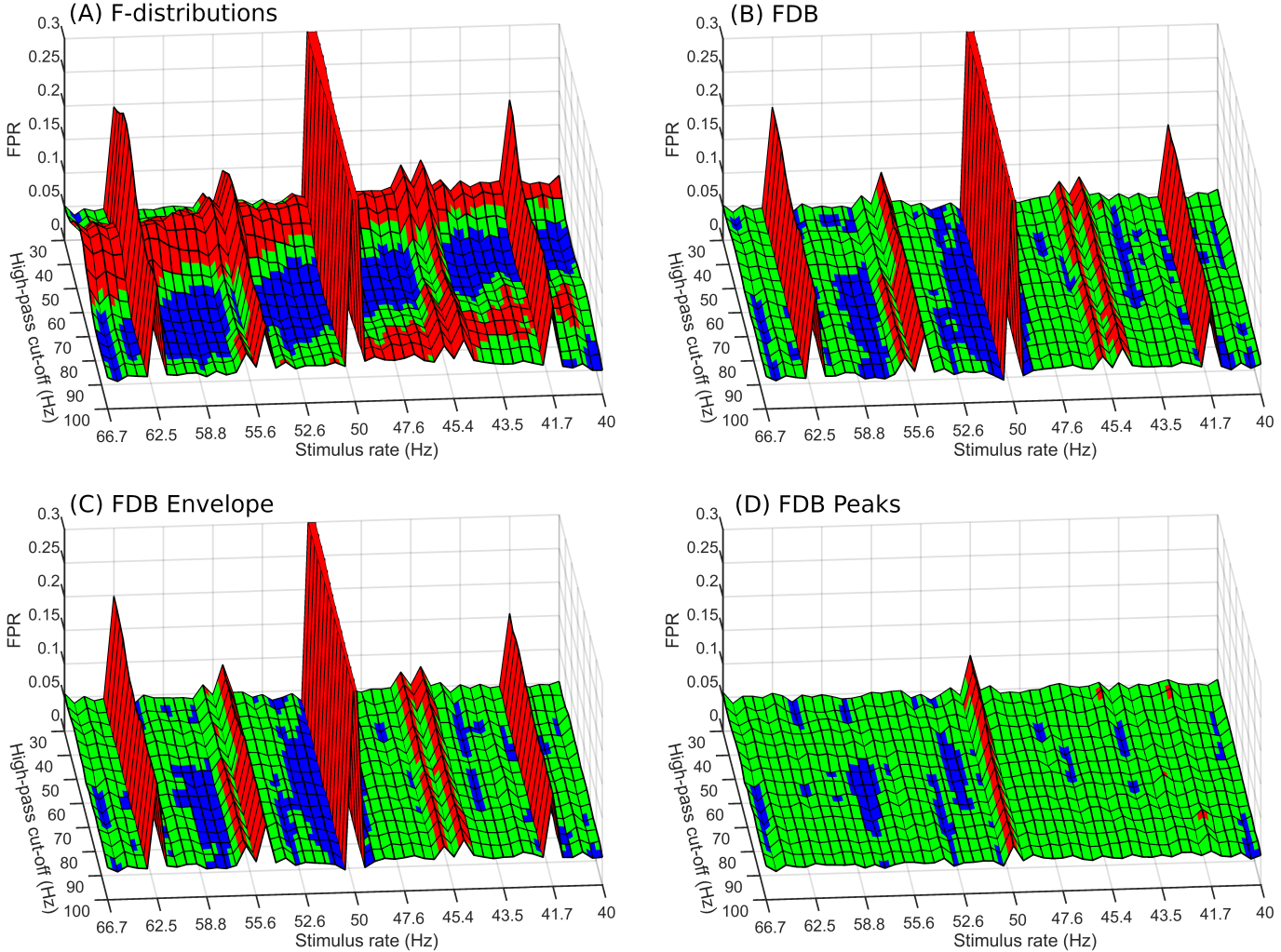
1. Power line noise

With respect to the large FPR peak at 50 Hz (reaching an FPR of ~0.4) in panels (A) - (C), this can be attributed to the 50 Hz power lines noise being time-locked to the epochs, resulting in residual power from the mains in the coherently averaged epoch. The additional peaks at e.g. 62.5 Hz and 41.67 Hz are likely due to the harmonics of the mains. To clarify this with an example: 6 periods of the 250 Hz harmonic gives a duration of 6x4ms = 24 ms, which coincides with one of the large FPR peaks at a stimulus rate of 1/0.024 = 41.667 Hz. Similarly, 4 periods of the 250 Hz harmonics gives a duration of 4x4ms = 16 ms, which corresponds to the 62.5 Hz peak. At 50Hz (and some other frequencies), multiple harmonics contribute to the increased FPR.

It can also be seen that the power lines noise has impacted on the FPRs for the frequencies surrounding the 50 Hz mains and its harmonics, albeit when using the original FDB (panel B) or the “FDB Envelope” modification (panels C). This can be attributed to the smoothing effect of the sliding Hanning window, as discussed previously. The latter was overcome with the FDB modification “FDB Peaks”, as peaks in the PSD were taken into account during the PSD estimation procedure.

1. Heteroskedasticity

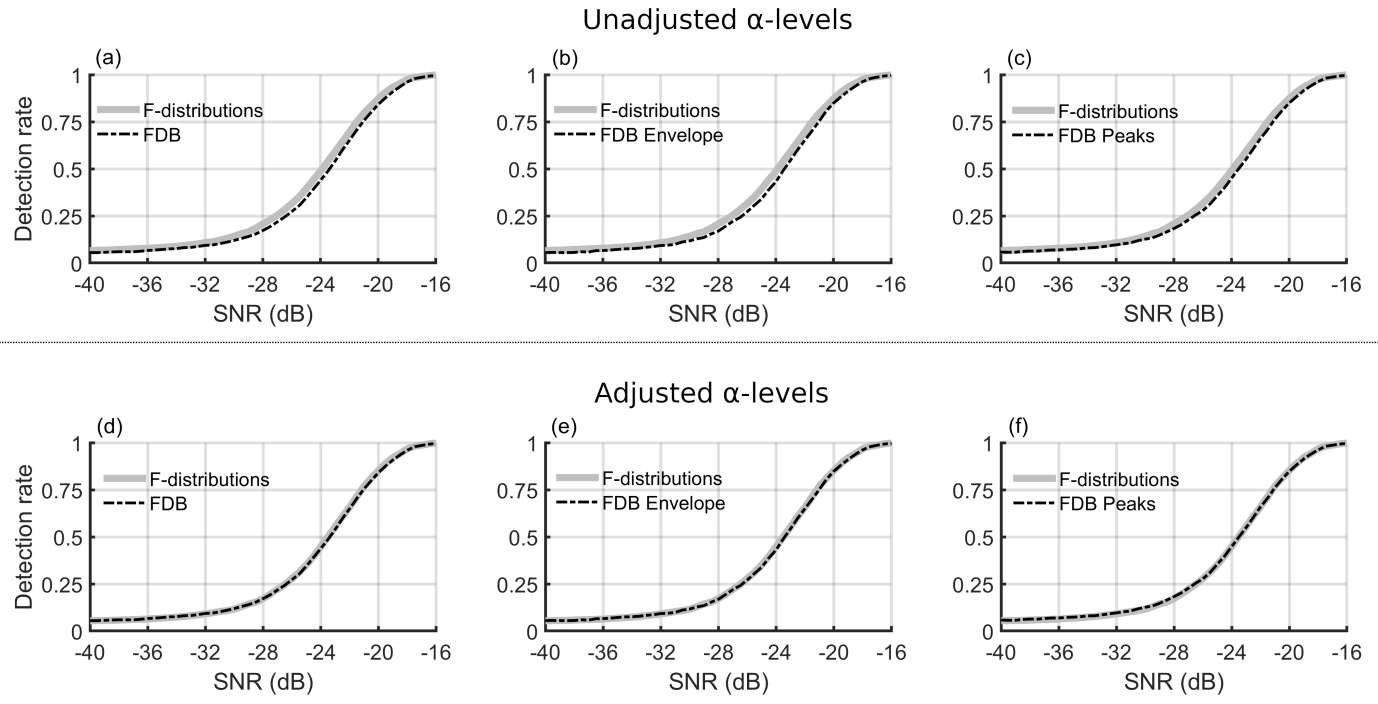
The discrepancy in FPRs between the conventional FDB (panel B) and the “FDB Envelope” approach (panel C) is relatively small, which suggests that non-stationary variance had a negligible impact on the T2 statistic for these recordings. Results obtained when simultaneously employing both modifications to the FDB were very similar to results shown in panel (D).



**Figure 4.** FPRs (α=0.05) for the subject-recorded no-stimulus data as a function of the high-pass cut-off frequency and (hypothetical) stimulus rate. Test significance was evaluated using (A) conventional F-distributions, (B) the original frequency domain bootstrap (FDB) approach, (C) the modified FDB approach “FDB Envelope”, or (D) the modified FDB approach “FDB Peaks”. FPRs shown in red indicate a significantly (p<0.01) liberal test performance (FPR > α=0.05), whereas FPRs shown in blue indicate a significantly conservative test performance (FPR < α=0.05) and FPRs in green fall within the 99% confidence intervals for α=0.05. The “wave-like” pattern in FPRs in panel (A) - starting from the upper left corner, moving towards the lower right - can be attributed to a violation of the independence assumption between epochs underlying the parametric HT2 test, whereas the large peaks at 62.5 Hz, 50 Hz, 41.67 Hz, etc., can be attributed to the mains interference and its harmonics.

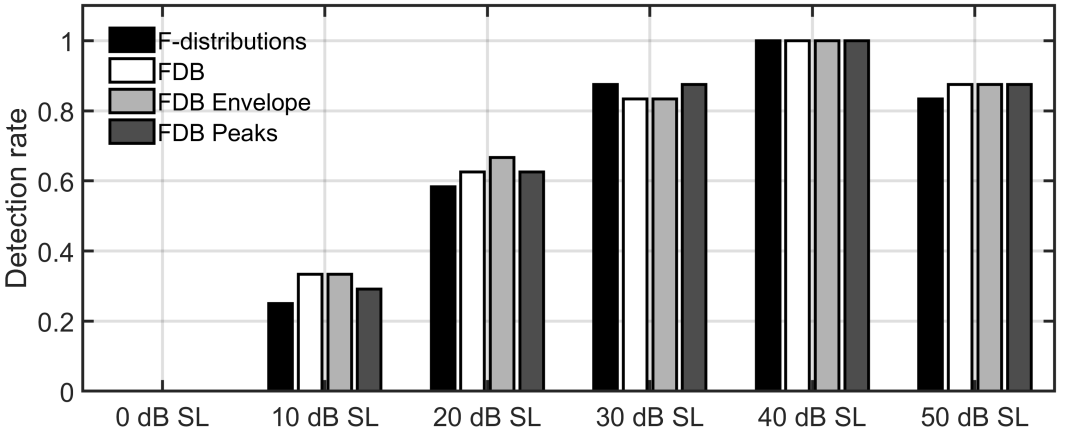
**3.2. Sensitivity**

**Simulations:** Detection rates (α=0.05) from the sensitivity assessment using simulations are plotted as a function of the SNR in panels (a), (b) and (c) in Figure 5. Results show slightly higher detection rates for the parametric approach (F-distributions) over the FDB and its modifications, which was attributed to a discrepancy in test specificity. After adjusting the α-levels so that all methods had equal FPRs, detection rates were more or less identical, as shown in panels (d), (e) and (f). The adjusted α-levels were 0.0402 for F-distributions and 0.0505 for the FDB and its modifications, giving FPRs of 0.0497, 0.0497, 0.0497, and 0.0494 for F-distributions, the FDB, the FDB Envelope approach, and the FDB Peaks approach, respectively. While the FDB approaches thus do not provide an increase in sensitivity (for the same specificity), they provide an improved control over the FPR without the need to adjust the α-levels, for which no-stimulus data was required.



**Figure 5.** Detection rates from the sensitivity assessment as a function of the simulated signal to noise ratio (SNR). Panels (a), (b), and (c) show the detection rates when using a fixed α-level of 0.05, whereas panels (d), (e) and (f) show the detection rates after adjusting the α-levels so that all methods had equal false-positive rates.

**Subject-recorded ABR threshold series:** Detection rates for the subject-recorded ABR threshold series are presented as bar plots in Figure 6, per dB SL condition. Visually inspecting the results suggests a similar test sensitivity for all methods. Fisher’s test also did not reveal significant differences between detection rates (p>0.05 for all pair-wise comparisons). The slightly lower detection rates for the 50 dB SL condition relative to the 40 dB SL condition can likely be attributed noise levels: for the 40 dB SL recordings, the mean power (averaged across recordings) was 8.6 uV, whereas for the 50 dB SL recordings this was 9.2 uV. This also resulted in more data being discarded due to artefact rejection, and hence in a less powerful test: Mean ensemble sizes were 1395 and 1212 epochs for the 40 and 50 dB SL conditions, respectively.



**Figure 6.** Detection rates for the subject-recorded ABR threshold series per dB SL condition. Test significance for the statistic was evaluated using either F-distributions, the frequency domain bootstrap (FDB), or the modified FDB approaches “FDB Envelope” and “FDB Peaks”.

**4. Discussion**

This work demonstrated the risk of liberal (FPRs > α) or conservative (FPRs < α) test performances for the parametric HT2 test when serial correlation was not taken into account (Figure 4). A liberal test is firstly undesirable, as false-positives can potentially have severe repercussions in some ABR applications, e.g. in new born hearing screening where false-positives can lead to undetected hearing, which can have a long-lasting impact (Ramkalawan and Davis, 1992; Yoshinaga-Itano et al. 1998; Idstad et al, 2019). A conservative test, however, is also undesirable, as this can result in lower than expected test sensitivities and prolonged test times, or even the need for further (repeated) testing with the associated costs and potential distress for patients and parents/carers. Ideally, the FPR should be controlled as intended, i.e. it should equal the pre-specified α-level of the test, as this facilitates test optimisation and helps to control errors.

The problem of serial correlation was overcome by adopting the FDB, which allowed the null distribution to be approximated under the sample correlation structure, ultimately providing an improved control over the FPR relative to the parametric approach. To obtain these results, it is important that the estimated PSD function  is an accurate approximation of the true PSD function of the EEG background activity. The accuracy of  is firstly impacted by its resolution, and hence the length of the sliding Hanning window. Additional simulations of EEG signals suggest that a 200 ms Hanning window (giving a 5 Hz frequency resolution), sliding across ~4 second recordings resulted in a small conservative bias in detection rates: FPRs ≈ 0.046 for α = 0.05, which was also observed in Figure 3. When the amount of data for estimating was increased to 12 seconds, the bias was reduced to ~0.049. The accuracy of  is also impacted by its variance, which can be reduced by increasing the amount of data, and/or by decreasing the length of the sliding Hanning window. Note that shorter windows increase the smoothness of , which will tend to give a good estimate of the general contour of , but may result in spectral details being distorted, including any sharp peaks in .

The attenuation and smoothing of peaks in  is undesirable, except when the peaks are induced by stimuli. In this case, attenuation is advantageous, as  will otherwise be biased towards the PSD function of the stimulus condition (EEG background activity + response), as opposed to the no-stimulus condition (EEG background activity only). This would unnecessarily increase the threshold for rejecting the null hypothesis, and reduce test sensitivity. Results from the sensitivity assessment nevertheless show that stimulus-induced peaks had a negligible impact on test sensitivity, albeit when using a 200 ms Hanning window (Figure 5). It is hypothesised that longer Hanning windows (and less smoothing) may result in a small decrease in test sensitivity, although this has not yet been tested.

**Accounting for peaks in the PSD function**

In order to prevent the attenuation and smearing of peaks in the  estimates, the “FDB Peaks” modification was introduced, which was shown to have a beneficial effect in simulations and subject-recorded data (Figures 3 and 4). It is envisaged that the FDB Peaks modification could also be used to circumvent stimulus-induced peaks in and prevent reduced test sensitivities when using longer windows. One limitation for the “FDB Peaks” approach worth emphasising is that peaks need to be identified in advance, prior to estimating , which might not be feasible for all noise sources. One noise source that can generally be predicted is the power lines noise, which was assumed to lie within ±0.5 Hz intervals around the 50 Hz and its harmonics.

Some alternative FDB methods for generating surrogate data with potentially non-smooth PSD functions include the local FDB (Paparoditis & Politis, 1999) and the auto-regressive aided FDB (Kreiss & Paparoditis, 2003). The auto-regressive aided FDB is essentially the regular FDB with an additional “pre-whitening” stage, and consists of first fitting an auto-regressive (AR) model to the data, and then applying the regular FDB to the residuals of the AR fit. Peaks in the PSD function are then potentially captured by the AR model, after which any non-white structure in the residuals is captured by the regular FDB. Whether this approach is able to estimate the sharp peaks in the PSD (due to e.g. power lines noise) without using a high AR model order (and potentially overfitting to the data ) is, however, still questionable, and has not yet been tested. An additional concern is that peaks induced by the stimuli might also be captured by the AR model, potentially biasing the surrogates towards the stimulus condition and reducing test sensitivity.

With respect to the local FDB (Paparoditis & Politis, 1999), this approach generates surrogate periodograms by applying bootstrap resampling within a sliding window, sliding across the periodogram ordinates. The assumption is that the spectrum is flat over a small range of adjacent frequencies. Note that this approach does not require the PSD function to be estimated, as bootstrap resampling is applied directly to the periodogram ordinates. A potential drawback is that a relatively high resolution of the sample perdiodogram and/or a short sliding window length is required, else the assumption that the spectrum is approximately flat within the sliding window might be violated. Stimulus-induced peaks should ideally also be excluded from the resampling procedure, else the surrogates might again be biased towards the stimulus condition. This approach also has not yet been evaluated for AEP detection.

**Accounting for heteroskedasticity**

The stationarity assumption underlying the FDB is a potential concern for ABR data, as some noise sources may be intermittent. The FDB Envelope approach was therefore introduced, which aims to reduce the impact of non-stationarities by rescaling the recording with its RMS envelope prior to PSD estimation. This approach was shown to be beneficial in simulations when data was clearly hetereoskedastic, but had little impact in the subject-recorded no-stimulus data used in the current work, which suggests that heteroskedasticity in the ABR data was relatively minor. However, this may not be the case in all experimental or recording conditions.

The main assumption underlying the FDB Envelope modification is that the estimated envelope is anaccurate estimate of the RMS envelope of the EEG background activity, which is impacted by the length of the sliding window used in its estimation: when this window is too long, will be ill-suited at capturing short-term changes in RMS amplitudes, whereas when it is too short, the RMS envelope may show large random fluctuations. The optimal length will depend on the specific application and could fail if there are strong “bursty” artefacts, such as eye-blinks in cortical evoked potentials.

Alternative methods for modelling correlated time-series with non-stationary variance are autoregressive conditional heteroskedasticity (ARCH) and generalised ARCH (GARCH) models (Francq & Zakoian, 2019). These methods first fit a parametric model to the data, after which the variance of the residuals is explained with an additional model, built around conditional variance, i.e. variance at time *t* is modelled as a function of the variance at previous time points. Bootstrap resampling in ARCH and GARCH models has previously also been considered by Hall & Yao (2003). Another option might be to substitute the AR model in the “pre-whitening stage” in the AR-aided FDB (Kreiss & Paparoditis, 2003) with an ARCH or GARCH model. These methods might be better suited at emulating rapid changes in variance, but have, to the best of the authors knowledge, not yet been tested for AEP detection.

**Other types of non-stationarity and block-wise application**

Besides heteroskedasticity, data can potentially also be non-stationary due to changes in the recording mean and/or auto-covariance function. Non-stationarities in the mean are generally not a concern for ABR data, as the recordings are high-pass filtered. Non-stationarities in the auto-covariance function, however, could again lead to unpredictable FPRs. The FDB method could be adjusted for this case by applying it in blocks, in which case the recording would be represented by multiple PSDs, as oppoted to a single PSD. A more sophisticated approach that also optimises how data is partitioned in blocks (albeit using an AR model) can be found in Davis et al (2006).

Another general approach to generate non-stationary surrogate data is constrained randomisation (Schreiber, 1998). This algorithm requires the desired data characteristics for the surrogates to be quantified in a cost function, after which the cost function is minimised (using e.g. the simulated annealing algorithm) for all possible permutations of the data. The approach is highly flexible and can impose a wide range of data characteristics on the surrogates, but comes at a relatively steep cost in terms of computational complexity.

There are a large number of alternative methods available for generating non-stationary surrogate data - an exhaustive review is outside the scope of this work. Some of the more well-known methods include the local block bootstrap (Paparoditis E., Politis, 2002) and various methods built around wavelets (e.g. Breakspear et al, 2003; Keylock, 2006; Keylock, 2007). A great overview on methods for generating both stationary and non-stationary surrogate data, along with some of their pros and cons, can be found in Lancaster et al (2018).

**Power lines noise and notch filters**

The large FPR peaks associated with the 50 Hz power lines noise and harmonics in Figure 4 raise the question as to whether including a notch filter in the pre-processing stage might have been beneficial. The specificity assessment for the HT2 test (evaluated with F-distributions) was therefore repeated using a 50 Hz notch filter (Lightfoot et al, 2014), and again with multiple notch filters, applied at 50 Hz mains and all harmonics, up to 1500 Hz. Results (presented in the Appendix) show that a single notch filter at 50 Hz did not prevent the large FPR peak at 50 Hz, and had an adverse effect on test specificity for the surrounding stimulus rates (within the ~50 ± 5 Hz region). When using multiple notch filters, the FPR peaks were prevented, but many new FPR peaks were introduced, along with an overall bias that was generally liberal (FPR > α) across all pre-processing parameters.

**How many surrogates?**

When approximating null distributions for statistical tests, many surrogates are needed to obtain a sufficiently smooth distribution. How many exactly depends on the application in question along with the desired accuracy for the pvalue. Firstly, note that the discrete nature of the bootstrapped distribution implies that it cannot distinguish between pvalues with differences smaller than 1/Bwhere Bis the number of surrogates. Secondly, the random resampling procedure underlying the FDB introduces an additional error term, which reflects the extent to which the bootstrapped distribution randomly deviates from its asymptotic distribution, achieved for . Note also that empirical observations become increasingly rare when approaching the extreme ends of the distribution. A large B is therefore particularly important if a high accuracy is desired for very small (e.g. < 0.001) or very large (e.g. >0.999) pvalues. A general rule of thumb for practical applications is to use  (Zoubir, 2004). For additional guidelines on choosing B, the reader is referred to Hall & Titterington (1989) or Zoubir (2004).

**Applications**

The main goal for the FDB and FDB modifications in this work was to obtain a good control over the FPR for ABR detection across a wide range of test conditions. By using bootstrap rather than theory-derived sampling distributions to calculate p values, the method is very flexible in allowing a wide range of statistical detection methods to be employed – including ones that are mathematically intractable.

An additional use for the FDB is that it may allow detection methods to be integrated with more efficient pre-processing and denoising methods where independence violations are more severe, e.g. when using faster stimulus rates or advanced denoising methods. Various denoising methods from image processing, for example, have shown promising results for improving the SNR of evoked responses ([Strauss](https://ieeexplore.ieee.org/author/37283368700) et al., 2002; Huang et al., 2015) but introduce additional correlation structures when doing so, which need to be taken into account when applying statistical detection methods.

Finally, the FDB might find applications in machine learning methods such as deep neural networks, which have recently also made their way to AER detection (McKearney & MacKinnon, 2019). In particular, the FDB could be used to augment existing data sets with large amounts of realistic background activity, which may help with exploring and optimising different algorithms, or offer new opportunities for transfer learning, e.g. deep neural networks might be pre-trained using FDB-generated surrogate data, after which the network could be ‘tweaked’ using recorded EEG data.

**Limitations**

A first limitation is that the FDB and FDB modifications were evaluated for a single test statistic (the statistic). Although the FDB procedure is expected to work well for other test parameters also (such as the Fsp, Fmp, Q-samples test, etc.), it may need to be further adjusted to obtain an optimal test performance. It is also worth noting that for frequency domain detection methods, the extent to which the underlying statistical assumptions are violated will also depend on which spectral bands are included in the analysis.

A second limitation is that non-stationarities in the autocovariance function, and their impact on the FDB and its modifications, were not systematically evaluated in the current work. It is worth noting that similar FPRs were observed when repeating the specificity assessment for ~16 second recordings (as opposed to ~4 seconds), which suggests that non-stationarities had a relatively minor impact for longer recordings also. While the use of successive 4s blocks was proposed as a means of dealing with nonstationarity in autocorrelation (or power spectrum), this was not further evaluated.

While statistical detection methods that are based on assumptions allow the sampling distributions to be determined *a priori*, the FDB methods requires the simulated data to be generated for each recording in order to take the recording-specific signal characteristics into account. This leads to a computational overhead and time-delay in obtaining results. For the ~4s ABR recordings in this study, the FDB method took ~2.7s to generate and analyse 5,000 surrogates using Matlab® code running on a fairly standard Windows Laptop (Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz) without careful optimization. There is much scope for improving this with more carefully developed software and more powerful hardware.

Finally, the current work evaluated test performance for ABR detection in normal hearing adults only. In future work, the FDB and its modifications could be tested using a larger cohort of test subjects, potentially with a range of hearing impairments, and/or in other evoked potential modalities.

**5. Conclusion**

This work demonstrated the risk of inflated or conservative FPRs for the conventional parametric HT2 test, which was attributed to serial correlation. The problem of serial correlation was overcome with the FDB, which allowed the null distribution to be approximated under the data correlation structure, ultimately resulting in an improved control over the FPR. A potential limitation for the conventional FDB, however, is that data is assumed to be stationary with a smooth PSD function. Two FDB modifications were therefore proposed, which aim to account for heteroskedasticity (using the “FDB Envelope” modification) and non-smooth PSD functions (the “FDB Peaks” modification). The advantage for the FDB Envelope approach over the conventional FDB in terms of specificity was demonstrated in simulations, but it had little impact in the subject-recorded data, which suggests that the impact of heteroskedasticity on the T2 statistic was negligible for this data set. The advantage for the FDB peaks approach over the conventional FDB was demonstrated in both simulations and in subject-recorded data and was due to an improved estimate of the underlying PSD function in the presence of peaks. In terms of test sensitivity, all methods showed similar detection rates in both simulations and subject-recorded data, for equivalent specificity. Overall, results suggest that the FDB and FDB modifications provided accurate, recording-dependent approximations of null distributions, and resulted in a good control over the FPR across a wide range of test parameters.

**Declarations of interest**

None

**Author contributions**

**Chesnaye M.A:** Conceptualisation, methodology, software, formal analysis, writing - original draft preparation. **Bell S.L.:** Conceptualisation, methodology, supervision, writing - reviewing and editing. **Harte J.M.:** Conceptualisation, methodology, writing - reviewing and editing. **Simpson D.M.:** Conceptualisation, methodology, supervision, writing - reviewing and editing.

**Acknowledgements**

The authors would like to acknowledge Debbie Cane and Sarah M.K. Madsen for collecting the subject recorded ABR data and the recordings of EEG background noise, respectively. The use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, were used in the completion of this work. This work was supported by the Oticon Foundation.

**6. References**

Breakspear M., Brammer M., Robinson P.A. Construction of multivariate surrogate sets from nonlinear data using the wavelet transform. 2003. Physica D, 182(1), pp. 1-22. doi: 10.1016/S0167-2789(03)00136-2

Brockwell, P., & Davis, R. A. (1991). Time series: theory and methods (2nd ed.) New York: Springer.

Chesnaye, M. A., S. L. Bell, J. M. Harte, and D. M. Simpson. 2018. “Objective Measures for Detecting the Auditory Brainstem Response: Comparisons of Specificity, Sensitivity, and Detection Time.” International Journal of Audiology 57 (6): 468–478. doi: 10.1080/14992027.2018.1447697.

Chesnaye M.A., Bell S.L., Harte J.M., & Simpson D.M. 2019. A group sequential test for ABR detection. Int. J. Audiol. doi: 10.1080/14992027.2019.1625486.

Cebulla M., Stürzebecher E., Elberling C. 2006. Objective detection of Auditory Steady State 709 Responses: Comparison of One-Sample and q-Sample Tests. J Am Acad Audiol, 17, 93-103. doi: 10.3766/jaaa.17.2.3

Dahlhaus R & Janas D. 1996. A frequency domain bootstrap for ratio statistics in time series analysis. Ann. Statist. 24(5): 1934-1963. doi: 10.1214/aos/1069362304

Dobie, R.A., Wilson, M.J., 1989. Analysis of auditory evoked potentials by magnitude-squared coherence. Ear Hear., Vol. 10(1), p.2–13. doi: 10.1097/00003446-198902000-00002

Duncan, C.C., Barry, R.J., Connolly, J.F., et al. (2009). Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. Clinical Neurophysiolog, 120(11), 1883-1908. doi: 10.1016/j.clinph.2009.07.045

Efron B., & Tibshirani R.J. (1993). An Introduction to the Bootstrap. Chapman & Hall/CRC.

Elberling C. & Don M. (1984). Quality estimation of averaged auditory brainstem responses. Scandinavian audiology, 13(3), pp. 187-197. doi: 10.3109/01050398409043059

Fisher R.A. Statistical methods for research workers, 11th ed. Oliverand Boyd, Edinburgh, 1932. doi: 10.1007/978-1-4612-4380-9\_6

Francq C & Zakoian J-M. 2019. GARCH Models: Structure, Statistical Inference and Financial Applications, Second Edition. John Wiley & Sons Ltd. doi: 10.1002/9781119313472

Geisler C.D. 1960. Average responses to clicks in man recorded by scalp electrodes. Massachusetts Institute of Technology, Research Laboratory of Electronics. Technical report 380

Golding M., Dilon H., Seymour J. & Carter L. (2009). The detection of adult cortical auditory evoked potentials (CAEPs) using an automated statistic and visual detection. International Journal of Audiology, 48(12), pp. 833-842. doi: 10.3109/14992020903140928

Hall P., & Titterington D.M. 1989. The Effect of Simulation Order on Level Accuracy and Power of Monte Carlo Tests. Journal of the Royal Statistical Society, Series B, 51: 459–467. doi: 10.1111/j.2517-6161.1989.tb01440.x

Hall, P., & Yao, Q. 2003. Inference in ARCH and GARCH models with heavy-tailed errors. Econometrica, 71, 285–317. doi: 10.1111/1468-0262.00396

Hotelling H. 1931. The Generalization of Student's Ratio. Ann. Math. Statist., 2(3), pp. 360-378. doi: 10.1007/978-1-4612-0919-5\_4

Huang Y., Chen X., Zhang J., Zeng D., Zhang D., & Ding X. 2015. Single-trial ERPs denoising via collaborative filtering on ERPs images. Neurocomputing, Vol 149, part B, pp. 914-923. doi: 10.1016/j.neucom.2014.07.043

Idstad M., Tambs K., Aarhus L., & Engdahl B.L. 2019. Childhood sensorineural hearing loss and adult mental health up to 43 years later:results from the HUNT study. BMC Public Health, Vol 19(168), doi: 10.1186/s12889-019-6449-2

Keylock C.J. Constrained surrogate time series with preservation of the mean and variance structure Phys. Rev. E, 73 (3) (2006), Article 036707. doi: 10.1103/PhysRevE.73.036707

Keylock C.J. A wavelet-based method for surrogate data generation. Physica D, 225 (2) (2007), pp. 219-228. doi: 10.1016/j.physd.2006.10.012

Kirch, C., & Politis, D. (2011). TFT-BOOTSTRAP: RESAMPLING TIME SERIES IN THE FREQUENCY DOMAIN TO OBTAIN REPLICATES IN THE TIME DOMAIN. The Annals of Statistics, 39(3), 1427-1470. doi: 10.1214/10-AOS868

Kreiss, J.-P., & Paparoditis, E. 2003. Autoregressive aided periodogram bootstrap for time series. The Annals of Statistics, 31, 1923–1955. doi: 10.1214/aos/1074290332

Kreiss J-P & Paparoditis E. 2011. Bootstrap methods for dependent data: A review. Journal of the Korean Statistical Society, 40, 357–378. doi: 10.1016/j.jkss.2011.08.009

Lahiri S.N. (2003) . Resampling Methods for Dependent Data. Springer Series in Statistics. Springer, New York, NY. doi: 10.1007/978-1-4757-3803-2\_9

Lancaster G., Iatsenko D., Pidde A., Ticcinelli V., & Stefanovska A. 2018. Surrogate data for hypothesis testing of physical systems. Physics Reports, Vol 748, pp.1-60. doi: 10.1016/j.physrep.2018.06.001.

Leemis L.M., & McQuestion J.L. 2008. Univariate Distribution Relationships. The American Statistician 62(February):45-53. doi: 10.1198/000313008X270448

Lightfoot G., Ferm I., Hall A., & Evans K. 2014. The effect of notch filtering on the waveform of the newborn auditory brainstem response, International Journal of Audiology, 53:9, 629-632, doi: 10.3109/14992027.2014.894644

Lightfoot G., Brennan S., FitzGerald J., & Ferm I. 2019. Recommended Procedure, Auditory Brainstem Response (ABR) Testing in Babies. The British Society of Audiology. Freely available online at https://www.thebsa.org.uk.

Lv J., Simpson D.M. & Bell S.L. (2007). Objective detection of evoked potentials using a bootstrap technique. Medical Engineering & Physics, 29(2), pp. 191-198.

doi: 10.1016/j.medengphy.2006.03.001

Madsen S.M.K., Harte J.M., Elberling C. & Dau T. (2018). Accuracy of averaged auditory evoked potential amplitude and latency estimates. International Journal of Audiology, 57(5), pp. 345-353. doi: 10.1080/14992027.2017.1381770

Mast J. & Victor J.D. (1991). Fluctuations of steady-state VEPs: interaction of driven evoked potentials and the EEG. Electroenceph. clin. Neurophysiol., 78(5), pp. 389-401. doi: 10.1016/0013-4694(91)90100-I

Maris E. & Oostenveld R. (2007). Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods, 164(1), pp. 177-190. doi: 10.1016/j.jneumeth.2007.03.024

Marple, S. L. Jr. 1987. Digital Spectral Analysis with Applications. Englewood Cliffs, NJ: Prentice-Hall.

McKearney R.M., & MacKinnon R.C. 2019. Objective auditory brainstem response classification using machine learning. Int. J. Audiol., 58, pp. 224-230. doi: 10.1080/14992027.2018.1551633

Møller AR. Evoked potentials in intraoperative monitoring. Baltimore: Williams & Wilkins, 1988.

Neely S.T. & Pepe M.S. (1997). United States Patent Number: 5,637,379. Method and Apparatus for Objective and Automated Analysis of Auditory Brainstem Response to Determine Hearing Capacity.

Paparoditis, E. (2002). Frequency domain bootstrap for time series. In H. Dehling, T. Mikosch, & M. Sörensen (Eds.), Empirical process techniques for dependent data (pp. 365–381). Boston: Birkhäuser. doi: 10.1007/978-1-4612-0099-4\_14

Paparoditis, E. E., & Politis, D. N. (1999). The local bootstrap for periodogram statistics. Journal of Time Series Analysis. Vol 20(2), pp. 192-222. doi: 10.1111/1467-9892.00133

Paparoditis E., Politis D.N. 2002. Local block bootstrap. Comptes Rendus Mathematique, Vol 335(11), pp.959-962. doi: 10.1016/S1631-073X(02)02578-5.

Picton, T.W. (2011). Human auditory evoked potentials. Chapters 11 & 12. Plural Publishing, San Diego, Oxford, Brisbane.

Pritchard, W. 1992.“The Brain in Fractal Time: 1/f-like Power Spectrum Scaling of the Human Electroencephalogram.” International JournalNeuroscience66 (1–2): 119–129.

doi: 10.3109/00207459208999796.

Punch, S., Van Dun, B., King A., Carter L., & Pearce W. 2016. Clinical Experience of Using Cortical Auditory Evoked Potentials in the Treatment of Infant Hearing Loss in Australia. Semin Hear., Vol 37(1), pp. 36–52. doi: 10.1055/s-0035-1570331

Ramkalawan T.W, & Davis A.C. 1992. The effects of hearing loss and age of intervention on some language metrics in young hearing-impaired children. British journal of audiology, 26(2), pp. 97-107. doi: 10.3109/03005369209077877

Rencher A.C. Methods of Multivariate Analysis. Second Edition. John Wiley & Sons, Inc., 2001

Schreiber T, Constrained randomization of time series data, Phys. Rev. Lett. 80, 2105 (1998).

Sininger Y.S., Hunter L.L., Hayes D., Roush P.A., & Uhler K.M. 2018. Evaluation of Speed and Accuracy of Next-Generation Auditory Steady State Response and Auditory Brainstem Response Audiometry in Children With Normal Hearing and Hearing Loss. Ear and Hearing, 39(6), pp. 1207-1223. doi: 10.1097/AUD.0000000000000580

Strauss D.J., Teuber T., Steidl G., & Corona-Strauss F.I. 2013. Exploiting the Self-Similarity in ERP Images by Nonlocal Means for Single-Trial Denoising. IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol 21(4), pp. 576-583. doi: 10.1109/TNSRE.2012.2220568

Valderrama J.T., de la Torre A., Alvarez I., Segura J.C., Thornton A.R.D. & Sainz M. (2014). Automatic quality assessment and peak identification of auditory brainstem responses with fitted parametric peaks. Computer Methods and Programs in Biomedicine., 114(3), pp. 262-275. doi: 10.1016/j.cmpb.2014.02.015

Valdes J.L., Perez-Abalo M.C., Martin V., Savio G., Sierra C., Rodriguez E. & Lins O. (1997). Comparison of Statistical Indicators for the Automatic Detection of 80 Hz Auditory Steady State Responses. Ear & Hearing, 18(5), pp. 420-429. doi: 10.1097/00003446-199710000-00007.

Vanheusden F., Bell S.L., Chesnaye M.A., & Simpson D.M. 2018. Improved Detection of Vowel Envelope Frequency Following Responses Using Hotelling’s T2 Analysis. Ear and Hearing, 40(1), pp.116-127. doi: 10.1097/AUD.0000000000000598

Victor J.D. & Mast J. (1991). A new statistic for steady-state evoked potentials. Electroencephalography and clinical neurophysiology, 78(5), pp. 378-388. doi: 10.1016/0013-4694(91)90099-p

Welch, P. (1967). The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modifified Periodograms. IEEE Transactions on Audio and Electroacoustics, 15, 70-73. doi: 10.1109/TAU.1967.1161901

Yoshinaga-Itano C., Sedey A.L., Coulter D.K., & Mehl A.L. 1998. Language of early- and later-identified children with hearing loss. Pediatrics, Vol 102(5), pp. 1161-71. doi: 10.1542/peds.102.5.1161.

Zoubir A.M. 2004. Bootstrap Techniques for Signal Processing. Chapter 3 - Signal detection with the bootstrap. Cambridge University Press. Page 82. doi: 10.1017/CBO9780511536717.005