Biomedical Signal Processing and Control 83 (2023) 104676

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Optimising weighted averaging for auditory brainstem response detection

Richard M. McKearney [a,*], Steven L. Bell [a], Michael A. Chesnaye [a], David M. Simpson [a]

[a] Institute of Sound and Vibration Research, Faculty of Engineering and Physical Sciences, University of Southampton, Southampton, UK

ARTICLE INFO

Keywords:
Auditory brainstem response
Weighted averaging
Evoked potentials
EEG
Signal processing

ABSTRACT

The auditory brainstem response (ABR) is a clinical test used to evaluate hearing objectively. The aim of this study was to optimise weighted averaging for both residual noise reduction and also for objective ABR detection using the Fmp statistical test. Analyses were performed using no-stimulus EEG background activity recorded from 15 participants and simulated "response present" data (4,602 ensembles in total). Different approaches for estimating the variance of the noise within each block were compared, as was the effect of the number of recording epochs in each block when calculating and applying the weights. The "VAR Whole Block" method was found to be more effective than the "VAR MP" method at estimating the noise level, especially for smaller block sizes (2–10 epochs). Caution should be exerted when selecting recording parameters for use with weighted averaging as an inflation in the "response absent" Fmp statistic was observed using small block sizes (relative to unweighted averaging); this may be due to a bias in the Fmp statistic observed as a result of the combined effects of the finite Fmp analysis window length and the high-pass filter setting. Optimised weighted averaging was effective in reducing the mean residual noise level in the averaged waveform, leading to improved ABR detection. Further work is required to optimise the Fmp analysis window length, recording settings, and weighted averaging parameters in combination, using a large clinical dataset.

1. Introduction

The auditory brainstem response (ABR) is a well-established clinical test, providing clinicians with an objective method of assessing hearing. The ABR consequently constitutes an integral component of numerous national newborn hearing screening programs [1].

Statistical detection methods may be used to assist clinicians in detecting if a response is present. If we define the recorded electroencephalogram (EEG) data for analysis as a matrix **X**, structured as N rows of recording epochs by M columns of samples within the chosen analysis window per recording epoch [2], then the t^{th} sample of the unweighted coherent average is calculated as:

$$\bar{x}[t] = \frac{1}{N}\sum_{i=1}^{N} x_i[t] \tag{1}$$

where $x_i[t]$ denotes the t^{th} sample of the i^{th} recording epoch. The presence of a response in the average can be evaluated using statistical tests, including the Fsp [3] and its close relation, the Fmp [4]. Both methods produce an F-statistic relating to the signal-to-noise ratio (SNR) of the averaged waveform, which may be converted into a p-value, quantifying the probability of obtaining the given result (or greater) under the assumption that the null hypothesis (no response present in the averaged waveform) is true. This is achieved by calculating a ratio of the estimated variance of the evoked potential signal (the coherent average) to the estimated variance of the averaged background noise and obtaining a p value for the F-statistic from the associated theoretical F-distribution. Specifically, the Fmp is calculated as [4]:

$$Fmp = \frac{VAR(\bar{\mathbf{x}})}{\frac{1}{N}\left(\frac{1}{Q}\sum_{j=1}^{Q} VAR(\mathbf{sp}_j)\right)} \tag{2}$$

where $\bar{\mathbf{x}}$ is the coherently averaged waveform, Q is the number of chosen single point noise estimates (**sp**) to be included in the analysis, and VAR $(\cdot)$ denotes variance. Note that $VAR\left(\mathbf{sp}_j\right)$ is found by taking the variance down the j^{th} chosen column of data matrix **X**. As a variance-ratio test, the Fmp is expected to produce a test statistic which follows an F-distribution under the condition that the null hypothesis ("response absent") is true. The degrees of freedom (df) of the numerator and the denominator are furthermore expected to be $\nu_1 = M-1$, and $\nu_2 = N-1$, respectively, under the assumption that the data follow a Gaussian distribution and

* Corresponding author at: Institute of Sound and Vibration Research, Faculty of Engineering and Physical Sciences, University of Southampton, SO17 1BJ, UK.
E-mail address: rm1n16@soton.ac.uk (R.M. McKearney).

https://doi.org/10.1016/j.bspc.2023.104676
Received 15 August 2022; Received in revised form 19 January 2023; Accepted 4 February 2023
Available online 13 February 2023

there is independence between samples both in the numerator (the coherent average) and the denominator (samples in each of the chosen single point noise estimates) [3]. However, due to the predominance of low frequency power in EEG background activity, there is correlation between samples within any given recording epoch, which reduces the degrees of freedom of the numerator (ν_1) [3]. The value of ν_1 is difficult to estimate on a recording-by-recording basis, and in order to safeguard against inflated false-positive rates in statistical tests, a conservative value of $\nu_1 = 5$ has been recommended with the majority of background EEG activity having been found to be at this level or greater [3]. The conservative choice for ν_1 comes at the expense of a reduced statistical power relative to larger values of ν_1.

Unweighted averaging is known to provide the maximum likelihood estimate of the response signal when the background noise is Gaussian white noise [5]. It is known that noise levels in the recording vary over time e.g. due to changes in EEG background activity, myogenic activity, or environmental sources of noise [6]. Some epochs will therefore contain more noise than others and have a proportionally lower SNR. A sequitur of this is that coherent averaging, which affords equal weight to all epochs (unweighted averaging), is inefficient for non-stationary noise and that applying a weight inversely linked with the noise level within each recording epoch will improve the SNR of the coherent average; this process is known as weighted averaging [6,7].

Applying weighted averaging to ABR recordings was proposed to address the shortcomings of unweighted coherent averaging, with the aim of reducing the residual noise levels within the coherent average [6,7]. When applying weighted averaging, the weights will typically be calculated as being inversely proportional to some estimate of the noise level within the EEG recording [6]. These weights may be calculated for individual recording epochs, however, the accuracy of the noise level estimate is limited by the amount of information regarding the noise level contained in such a short data segment. Elberling & Wahlgreen [7] therefore proposed that the (single point) noise variance estimate be calculated from blocks of several recording epochs, with the weight then being applied to the whole block. Elberling & Wahlgreen [7] provide the following equation to calculate the t^{th} sample of the weighted coherent average ($\bar{\mathbf{x}}_w$):

$$\bar{\mathbf{x}}_w[t] = \left(\frac{\bar{\mathbf{x}}_1[t]}{V_1} + \frac{\bar{\mathbf{x}}_2[t]}{V_2} + \cdots \frac{\bar{\mathbf{x}}_n[t]}{V_n}\right) \bullet \frac{1}{T} \tag{3}$$

where $\bar{\mathbf{x}}_n[t]$ is the t^{th} sample of the unweighted coherent average of the recording epochs within the n^{th} block of recording epochs, V_n is the estimated noise variance of the n^{th} block [3], and n is the number of blocks of recording epochs. The variable T constrains the weights to sum to unity, calculated as the sum of the reciprocal of the n noise variance estimates [7]:

$$T = \frac{1}{V_1} + \frac{1}{V_2} + \cdots \frac{1}{V_n} \tag{4}$$

Applying weighting inversely proportional to the noise power (variance) in each block provides a linear minimum mean square error (MMSE) estimator [6,7], and maximum likelihood estimator of the evoked potential signal [5].

Using larger blocks of recording epochs has the advantage of allowing the noise level to be estimated more accurately. However, this does not account for any non-stationarity occurring within the block, with all epochs within the same block being allocated the same weight. There is therefore a trade-off when deciding on the value of the block size parameter, between improving the accuracy of the noise level estimate and faster adaptation of the weights as the noise level changes over time [8]. Don & Elberling [8] sought to optimise the block size parameter for residual noise reduction, evaluating block sizes of 32, 64, 128, and 256, and using up to 8 samples from each recording epoch in a block to estimate the noise levels. The block size of 32 epochs reduced the residual noise levels in the weighted coherent average most efficaciously and the possibility therefore remains for an even smaller block size to be yet more effective.

Another aspect of weighted averaging which may benefit from optimisation, and was addressed by this study, is how best to estimate the variance of the background noise. Finally, whilst the effects of weighted averaging on the F-statistics for ABR detection have been demonstrated on isolated recordings [3], a systematic evaluation of the effects of weighted averaging on statistical methods for ABR detection has not yet been presented in the literature. The primary aim of this study was therefore to optimise weighted averaging, in terms of residual noise level and ABR detection, by comparing noise estimation methods for calculating the weights. The second aim was to optimise the block size parameter. The final aim was to assess the impact of weighted averaging when using the Fmp statistical test to detect the ABR objectively.

2. Materials and methods

2.1. "Response absent" background EEG data

The spontaneous EEG data used in this study were previously recorded from 17 participants by Madsen et al. [9,10] under several recording conditions: asleep, lying still, blinking, and with head movement. No sound stimuli were delivered. Only the EEG recordings from the "asleep" and "lying still" conditions were used in the current study as these conditions best reflect those under which the ABR is recorded in clinical practice, representing approximately 6.5 h of recordings from 15 participants. Offline processing of these data included band-pass filtering from 30 to 1,500 Hz using a 3rd order Butterworth filter and downsampling the data to 5 kHz. These filter parameters were chosen as they reflect the recommendations of the British Society of Audiology [1]. Artefact rejection was applied with the threshold level set at ±25 μV. A relatively high artefact rejection level was chosen to allow more noise into the recordings so as to allow better observation of the effects of weighted averaging. The continuous EEG recordings were arranged into ensembles of 1,000 epochs each. Recording epochs were 30 ms in length and were spaced temporally from each other to emulate a stimulus rate of 33.3 Hz (albeit with no stimulus being delivered). In total, 2,301 "response absent" ensembles of 1,000 recording epochs each were produced.

2.2. "Response present" ABR data

The "response present" data used in this study were simulated by making a copy of the 2,301 "response absent" ensembles and adding an ABR template to every recording epoch of each ensemble. The ABR template was constructed from the database described by Chesnaye et al. [2] and Lv et al. [11] (available at: https://doi.org/10.5258/SOTON/D0168), which contained 33.3 Hz click-evoked ABR data from 12 normal-hearing adults (6 females and 6 males), filtered offline in the same manner as the background EEG data. The single "response present" ABR template used was a coherently averaged waveform recorded at 50 dB SL (sensation level – relative to the individual's hearing threshold) and scaled to have a peak-to-peak amplitude of 500 nV. The benefit of using simulations in this study is that the "true" ABR signal is known *a priori*, hence allowing the estimation error in the ABR waveform to be calculated.

2.3. Ethics

Ethical approval for secondary data analysis of the datasets used in this study was granted by the University of Southampton Research Ethics Committee.

2.4. Evaluation of the operating characteristics of the weighted Fmp statistic

In order to find the Fmp statistic (Equation (2)) with weighted averaging, a weighted ensemble is first defined as follows by adapting Equation (3):

$$\mathbf{X}_w = \begin{bmatrix} \mathbf{X}_1 \bullet \left(\frac{n}{V_1 \bullet T}\right) \\ \mathbf{X}_2 \bullet \left(\frac{n}{V_2 \bullet T}\right) \\ \vdots \\ \mathbf{X}_n \bullet \left(\frac{n}{V_n \bullet T}\right) \end{bmatrix} \quad (5)$$

where $\mathbf{X}_n$ is the sub-ensemble formed by the n^{th} block of $L = \frac{N}{n}$ recording epochs from $\mathbf{X}$. The (unweighted) coherent average of the weighted ensemble $\mathbf{X}_w$ produces the weighted average $\bar{\mathbf{x}}_w$ from Equation (3), which was substituted for the unweighted average $\bar{\mathbf{x}}$ in Equation (2) when calculating the weighted Fmp statistic. The multiple single point noise estimates in the Fmp denominator were also calculated from $\mathbf{X}_w$, rather than the original unweighted ensemble $\mathbf{X}$ when calculating the weighted Fmp statistic. The Fmp statistic was calculated over an analysis window of sample points contained within 1–15 ms of the recordings, using all available columns within the analysis window for the noise estimate such that $Q = M$. The Fmp statistic was calculated for all 4,602 ensembles, for each of the block sizes evaluated.

Detection performance was measured using the partial area under the receiver operating characteristic curve (ROC AUC) performance metric [12]. For this application, the clinical area of interest on the ROC curve is the portion of the curve on the left, corresponding to high specificity levels. The partial ROC AUC was calculated for the region from 95 to 100% specificity.

2.5. Noise level estimation methods

Two approaches for calculating V_k (for $k = 1,2,\cdots,n$) in Equation (5) were evaluated, i.e. estimating the variance of the noise within each block. These noise level estimates were derived from the unweighted ensemble $\mathbf{X}$ and used to produce the weighted ensemble $\mathbf{X}_w$. The first method, "VAR MP", is similar to the original noise level estimation method which Elberling & Wahlgreen [7] used to perform weighted averaging, except that multiple points are used instead of a single point ($Q = 1$). For each block of recording epochs, the mean of multiple single point estimates of the variance of the noise is calculated $\left(\frac{1}{Q}\sum_{j=1}^{Q} VAR\left(\mathbf{sp}_j\right)\right)$ as used in the denominator of the Fmp ratio (Equation (2)):

$$V_{k,(VAR\ MP)} = \frac{1}{Q(L-1)}\sum_{j=1}^{Q}\sum_{i=1}^{L}\left(\mathrm{x}_{k,i,j} - \bar{\mathrm{x}}_{k,*,j}\right)^2 \quad (6)$$

where $\mathrm{x}_{k,i,j}$ is the EEG sample from the i^{th} row of the j^{th} chosen column of the k^{th} block of recording epochs and $\bar{\mathrm{x}}_{k,*,j}$ is the mean value of the j^{th} chosen column of the k^{th} block. In this study all columns within the 1–15 ms analysis window were used in the multiple point noise estimate, meaning that $Q = M$.

The second approach to estimating the noise level within each block (V_k), the "VAR Whole Block" method, is calculated as the variance of all of the ($L \times M$) concatenated samples within the block:

$$V_{k,(Whole\ Block)} = \frac{1}{ML-1}\sum_{t=1}^{M}\sum_{i=1}^{L}\left(\mathrm{x}_{k,i,t} - \bar{\mathrm{x}}_{k,*,*}\right)^2 \quad (7)$$

where $\bar{\mathrm{x}}_{k,*,*}$ is the average of all the samples in the k^{th} block. This method has the advantage of making more efficient use of the available information within the block to estimate the noise level and provides an unbiased estimator of the variance of the EEG background activity under the "response absent" condition. However, when an ABR signal ($\mathbf{s}$) is present, the estimate will be biased [5] by an additional $VAR(\mathbf{s})$ term (assuming a deterministic evoked potential signal which is independent of the background noise), where $VAR(\mathbf{s})$ denotes the variance of the evoked potential response. For low-amplitude evoked responses, the bias term in the variance estimate is expected to be negligible [5], but for larger responses this could lead to sub-optimal weighting. It was hypothesised that using the "VAR Whole Block" method would allow the noise variance to be estimated more accurately (albeit with a bias), particularly for small block sizes where "VAR MP" is estimated from few points and hence subject to large random estimation errors.

2.6. Residual noise estimation and optimisation of block size

For each of the 4,602 ensembles, the weighted coherent average ($\bar{\mathbf{x}}_w$) and weighted ensemble ($\mathbf{X}_w$) were calculated using weights provided by either the "VAR Whole Block" or the "VAR MP" approach for estimating the noise level V_k (for $k = 1, 2, \ldots, n$). The residual background noise in the weighted average of the "response present" data, could then be calculated accurately by subtracting the known ABR template from the weighted average $\bar{\mathbf{x}}_w$ (subtraction not required for the "response absent" data). The residual noise level in the coherent average was quantified by the root mean square (RMS) value of this difference signal, within the analysis window of 1–15 ms. The assessment was repeated over a range of block sizes to determine which block size optimally reduced the residual noise level within the averaged waveform. The block sizes used were equal to all of the factors of the ensemble size of 1,000 epochs: 1, 2, 4, 5, 8, 10, 20, 25, 40, 50, 100, 125, 200, 250, 500, and 1,000 epochs. Note that using a block size of 1,000 (i.e. all) epochs equates to unweighted ensemble averaging. Also note that the "VAR MP" method (Equation (6)) cannot be applied to a block size of 1.

3. Results

3.1. A comparison of the two noise level estimation methods and optimisation of block size

Fig. 1 provides a comparison of the "Var MP" and "Var Whole Block" methods in terms of their efficacy at reducing residual noise levels within the weighted average.

Fig. 1 shows that weighted averaging can achieve notable improvement in residual noise levels compared to unweighted averaging. The "VAR Whole Block" method consistently led to lower mean and median residual noise levels when used for weighted averaging compared to the "VAR MP" method. As expected, for larger block-sizes the RMS error gradually converged to that of unweighted averaging. However, as the block size becomes smaller, the noise level estimates may become inaccurate, leading to an *increase* in residual noise levels, which can exceed that of unweighted coherent averaging. For the "VAR Whole Block" method, a lower block size led to a quite consistently lower median residual noise level (relative to unweighted averaging), however, the mean residual noise level increased above that of the baseline of unweighted averaging. Whilst the largest median reduction in residual noise in the averaged waveform was achieved using the smallest block size of 1 epoch with the "VAR Whole Block" method, the largest mean reduction in noise was achieved using 25 epochs-per-block. Improved reduction in residual noise within the coherent average is expected to translate into improved detection of the ABR using the Fmp statistical test.

A further simulation was performed to determine the effects of the SNR of the data on the accuracy of the noise level estimation method (see Figure, Supplemental Digital Content 1). The results showed that if the SNR of the "response present" continuous EEG data (before averaging) was less than ~ -16 dB, then the "VAR Whole Block" method was more effective or of equivalent effectiveness to the "VAR MP" method at

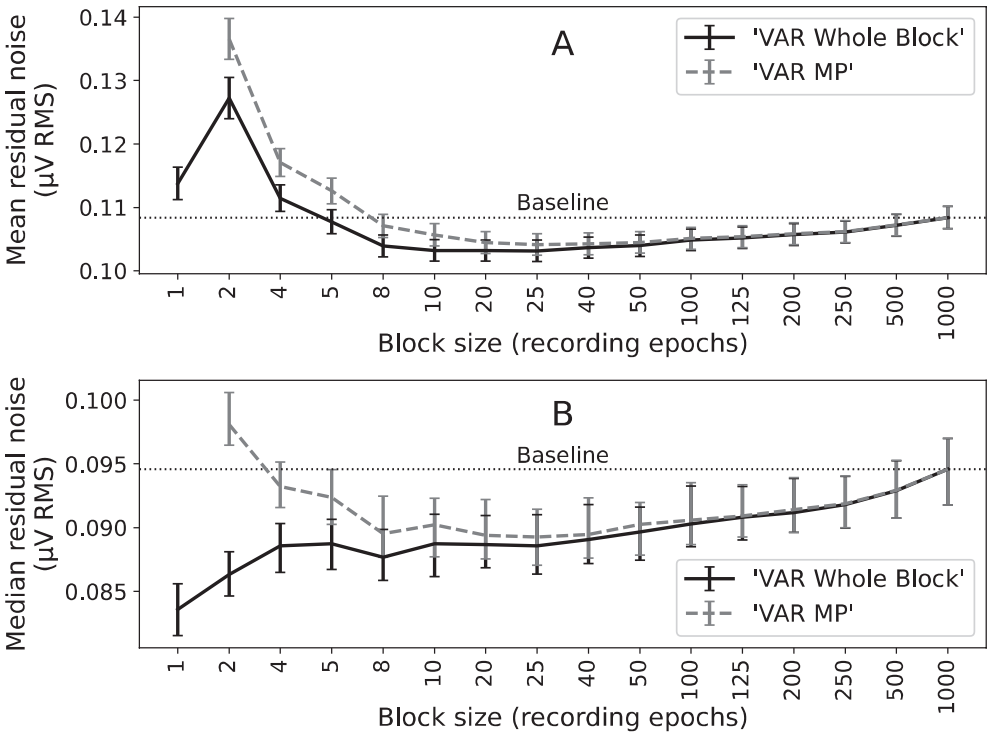

Fig. 1. The effects of the weighted averaging noise level estimation method on residual noise levels in the coherent average. The mean (Fig. 1A) and median (Fig. 1B) residual noise levels in the coherent average, are presented as a function of block size. Unweighted coherent averaging, obtained when using a single block (block size of 1,000), provided a baseline (dotted line) for comparison of performance. Note that a block size of 1 cannot be used for the "VAR MP" method. Error bars represent the 95% confidence interval of the mean (Fig. 1A) and the median (Fig. 1B) residual noise levels in the coherent average.

estimating the noise level, depending on the block size (the smaller the block size the larger the difference between the two methods). Where the SNR was above around −16 dB, the effectiveness of the "VAR Whole Block" method began to deteriorate substantially, as the bias present in the variance estimate increased.

3.2. The effects of weighted averaging on ABR detection using the Fmp

Fig. 2 provides a comparison of the performance of the Fmp in detecting the ABR when weighted averaging was carried out using either the "VAR MP" or "VAR Whole Block" noise level estimation method.

Fig. 2 shows that weighted averaging in conjunction with the Fmp criterion can lead to improved detection of responses (as measured by the partial ROC AUC), compared to unweighted averaging. A paired permutation test (using 5,000 permutations) was used to compare the "VAR Whole Block" method with the "VAR MP" method for each block size. The "VAR Whole Block" method (in combination with the Fmp) was able to achieve a statistically significantly higher partial ROC AUC than the "VAR MP" method, for block sizes between 2 and 10 (inclusive), indicating that this is a more powerful (combined) ABR detection method. The highest partial ROC score was achieved using the "VAR Whole Block" method and a block size of 2.

The previous analysis (Fig. 1) considered only the RMS error, which affects the numerator of the Fmp, however, block-wise analysis (weighted averaging) impacts upon both the numerator and the denominator. Whilst lower residual noise levels will aid detection, detection performance as measured using the partial ROC AUC is influenced by how well "response present" and "response absent" data can be separated based on their Fmp values. Further analysis was therefore undertaken providing separate evaluations for "response present" and

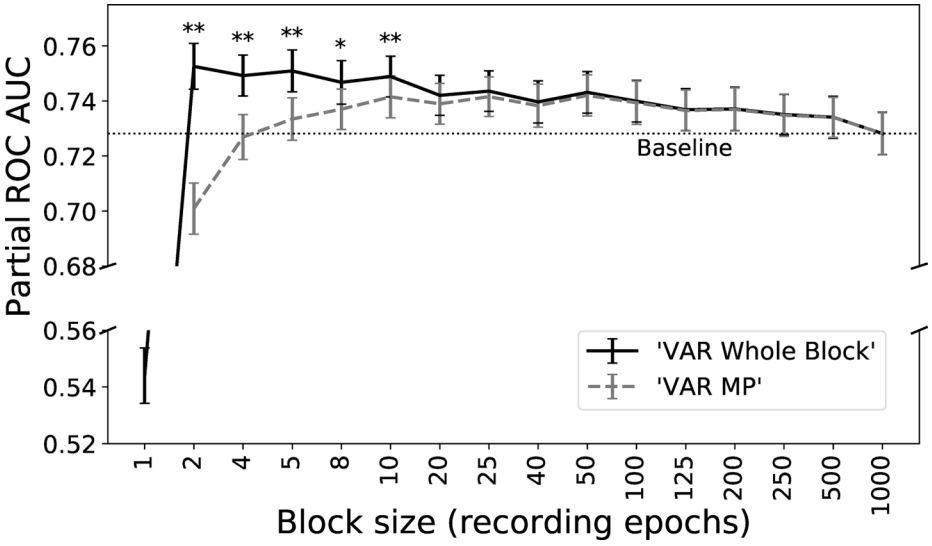

Fig. 2. Performance of the Fmp in detecting the ABR after weighted averaging. A comparison of the "VAR MP" and "VAR Whole Block" methods is presented based on the partial area under the receiver operating characteristic curve (ROC AUC) and the Fmp measure of the quality of the averaged ABR criterion. A higher partial ROC AUC corresponds to the detection method (Fmp combined with weighted averaging) having a better ability to discriminate between "response present" and "response absent" data for false positive rates of up to 0.05. The baseline value corresponds to the results from unweighted averaging. The error bars represent the bootstrapped standard error of the partial ROC AUC. A single asterisk, *, indicates a Bonferroni-adjusted two-sided p value of < 0.05. A double asterisk, **, indicates a Bonferroni-adjusted two-sided p value of < 0.01.

"response absent" data.

3.3. The effects of weighted averaging on the Fmp test statistic for "response present" and "response absent" data

Based on the results from Fig. 1 and Fig. 2, the more effective "VAR Whole Block" method was selected over the "VAR MP" method for use in the subsequently presented analyses. Fig. 3 shows the effects of weighted averaging (using the "VAR Whole Block" method) for both "response present" and "response absent" data.

Fig. 3 shows that smaller block sizes led to increased Fmp values (positive absolute change as opposed to negative), both for "response present" and "response absent" data. The plots also show that using smaller block sizes tended to lead to a greater increase in the Fmp value for the "response present" data (dashed grey lines, Fig. 3A and 3B). Of surprise, it was observed that the Fmp values for the "response absent" data began to increase when reducing the block size from the baseline of 1,000 epochs-per-block. Based on the assumption of the EEG background noise data comprising independent and identically distributed random variables following a Gaussian distribution, the expectation is that the Fmp value for "response absent" data would remain unchanged. The observed Fmp inflation for the "response absent" data may also lead to false positive rates that exceed the expected (nominal) values.

3.4. Investigating Fmp inflation in the "response absent" data

In order to investigate the Fmp inflation observed in the "response absent" data when weighted averaging was applied (Fig. 3), further analysis of the null distribution of the data were performed along with an analysis of the effects of weighted averaging based on the initial unweighted value of the Fmp statistic (Fig. 4).

Fig. 4B shows how the mean Fmp value of the "response absent" data changed with weighted averaging (using 2 epochs-per-block) as a function of the unweighted Fmp level. If an unweighted Fmp value was < 1, then weighted averaging tended to revert the Fmp value upwards towards 1. For ensembles with an unweighted Fmp value > 1, the reverse was true. As the mean Fmp value of the ensembles in the dataset was less than one, the mean Fmp was found to increase when applying weighted averaging (Fig. 3).

The empirically obtained mean unweighted Fmp value under the null condition (equal to 0.952 $\pm$ 0.020 [95% CI]) was found to be significantly lower than the expected value of an F-distributed random variable: $E[F] = \frac{\nu_2}{\nu_2 - 2} = 1.002$ [13]. A one-sample permutation test (using 20,000 permutations) was performed to test the hypothesis that there was no difference between the observed mean value of the null Fmp distribution and the expected value ($E[F]$), assuming that $\nu_2 = N - 1$ degrees of freedom; $p < 0.001$ (two-sided). This indicates that even the standard, unweighted Fmp value, as evaluated using the present dataset and analysis parameters, did not conform to the F-distribution expected from theory and with the standard assumptions. The reasons for this will be considered in the Discussion section. The effect of inflation of the Fmp test statistic in the null condition on the false positive rate is investigated in the next section.

3.5. Specificity analysis

Fig. 5 shows the specificity when using weighted averaging paired with the Fmp detection method with the critical value determined by the F-distribution with the assumed $\nu_1 = 5$ degrees of freedom [3]. The specificity levels achieved were significantly above those expected for block sizes of 2 to 1,000 (i.e. including unweighted averaging) as evidenced by the bars in Fig. 5 exceeding the binomial proportion 95% confidence interval (CI) for the nominal significance level of 0.01, calculated using the Wilson score interval method [14]. Whilst a low false positive rate is desirable, exceeding the target level of specificity comes at the expense of reduced sensitivity. The specificity level for 1 epoch-per-block was significantly below the expected 95% CI, likely due to the large inflation in the mean "response absent" Fmp value observed using this block size (Fig. 3). In order to control the false positive rate a bootstrap method for evoked potential recordings was evaluated [2,11].

3.6. Controlling the false positive rate using the bootstrap method

An alternative approach to calculating a p value for the Fmp statistic using the associated theoretical cumulative distribution function [3], is to apply a nonparametric bootstrap to the empirically obtained data [2,11,15,16]. For each ensemble being evaluated, the data are repeatedly resampled in a manner which disrupts the presence of any evoked

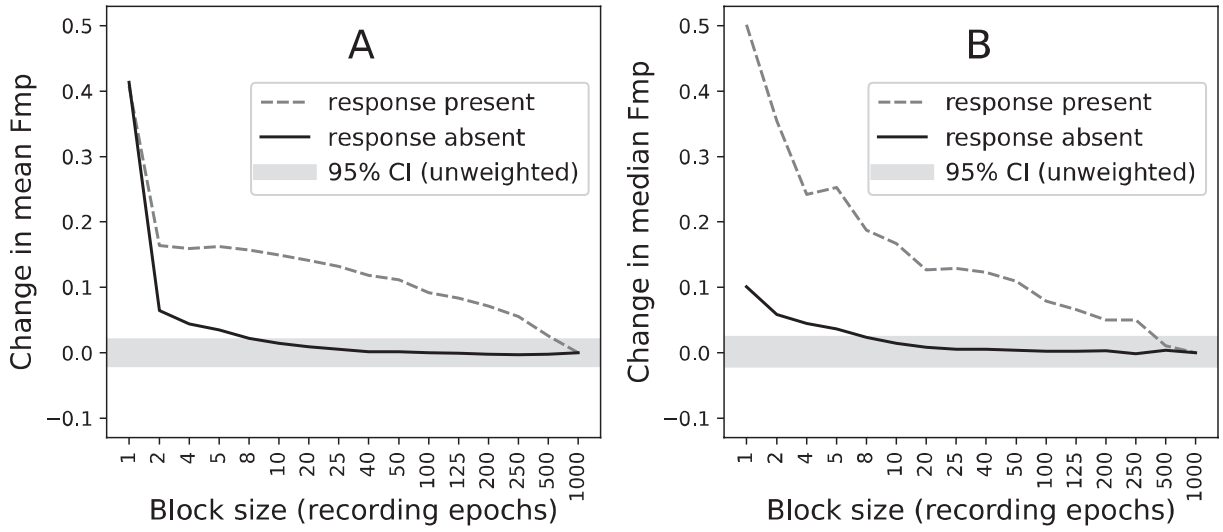

Fig. 3. The effects of weighted averaging on the Fmp test statistic. Here the data are presented as the absolute change in mean (Fig. 3A) and median (Fig. 3B) Fmp values. The absolute change was measured with reference to the baseline of the mean (Fig. 3A) or median (Fig. 3B) Fmp value of the data using unweighted averaging (1,000 epochs-per-block). The solid black and dashed grey lines correspond to "response absent" and "response present" data, respectively. The light grey band represents the 95% confidence interval for the mean (Fig. 3A) and the median (Fig. 3B) Fmp values of the "response absent" data in the unweighted condition. These confidence intervals may be used to evaluate whether the mean/median "response absent" Fmp values obtained using weighted averaging differ significantly to the mean/median obtained using unweighted averaging.

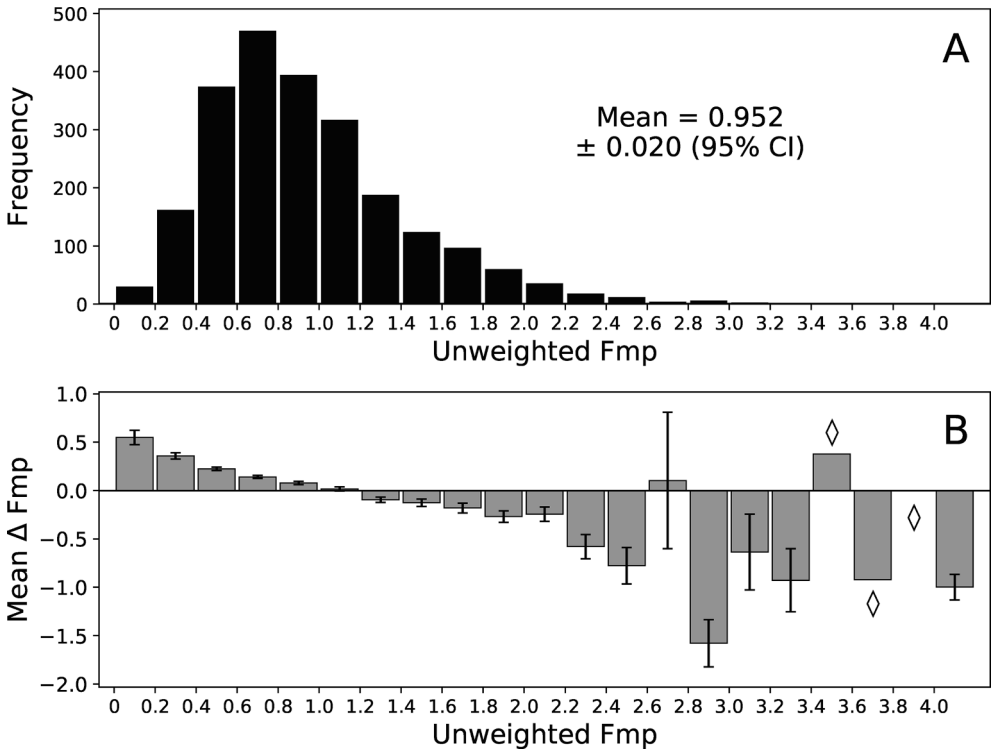

Fig. 4. The effects of weighted averaging on the "response absent" Fmp value. Fig. 4A shows the null distribution of the unweighted Fmp test statistic ("response absent" data). The mean Fmp value of this null distribution was 0.952 ± 0.020 (95% CI). When weighted averaging was applied the mean Fmp value increased from 0.952 to 1.017. Fig. 4B shows the mean change in Fmp value for the "response absent" data when applying weighted averaging (using 2 epochs-per-block), based on the initial unweighted Fmp value of the data. For ensembles with an initial unweighted Fmp value of <1, the mean Fmp value tended to increase when weighted averaging was applied. The further away from 1, the greater the effect was. For ensembles with an initial unweighted Fmp value of >1, the reverse was true. Note that for unweighted Fmp values in both the upper and lower extremes, there were fewer data points compared to the centrally occurring values leading to greater variability in results. Error bars represent the standard error (SE) of the mean (◊ denotes where the SE was not calculable due to a low sample size of ≤ 1).

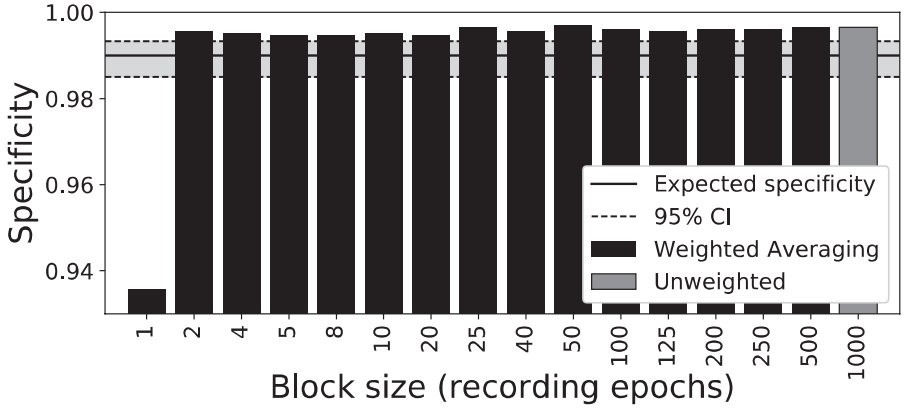

Fig. 5. Specificity measured across block sizes. The specificity level is shown as a function of the weighted averaging block size. A nominal significance level of 0.01 was selected based on the clinical need to have a high degree of confidence when declaring a waveform as containing a response [1]. This corresponds to a target specificity of 0.99. The binomial proportion 95% CI (grey), calculated using the Wilson score interval, was based on an expected false positive rate of 0.01 and n = 2,301 "response absent" trials.

potential signal in order to estimate the null distribution of the test statistic for that particular recording. A *p* value for the test statistic calculated from the original ensemble may then be obtained by observing the proportion of the bootstrapped null distribution that resides above that value [2,11]. The resampling is performed by selecting *N* random sections of *M* samples from the continuous EEG to form an ensemble. As these *N* recording epochs are selected randomly from anywhere within the continuous EEG, irrespective of their timing in relation to the stimulus onset, the evoked potential will be disrupted and hence reflect the null condition ("response absent") [2,11]. Through repeated resampling, the null distribution may be approximated; in the current study 500 bootstrap samples per ensemble were used. The results obtained from the set of 2,301 "response absent" ensembles show that the bootstrap was effective in controlling the false positive rate (Fig. 6A). Thus, by combining the weighted average, using the "VAR Whole Block" method to estimate the weights, and the Fmp combined with the bootstrap method, the ABR detection rate increased from the initial (unweighted) level of 0.415 to 0.500 using the smallest block size (Fig. 6B).

4. Discussion

Weighted averaging is an effective method for reducing the residual noise levels present in the coherently averaged waveform, placing greater emphasis on recording epochs with lower noise levels relative to those containing higher noise levels [6,7]. This study has sought to further optimise the weighted averaging procedure as well as provide an analysis of the effects of weighting averaging on automated ABR detection, focusing specifically on the Fmp statistical test. The results clearly show that weighted averaging can both improve the quality of the estimated ABR signal, and also its detection using the Fmp, when using appropriately selected parameters.

4.1. Observed bias in the Fmp statistic

In the present study, the empirically observed mean value of the Fmp

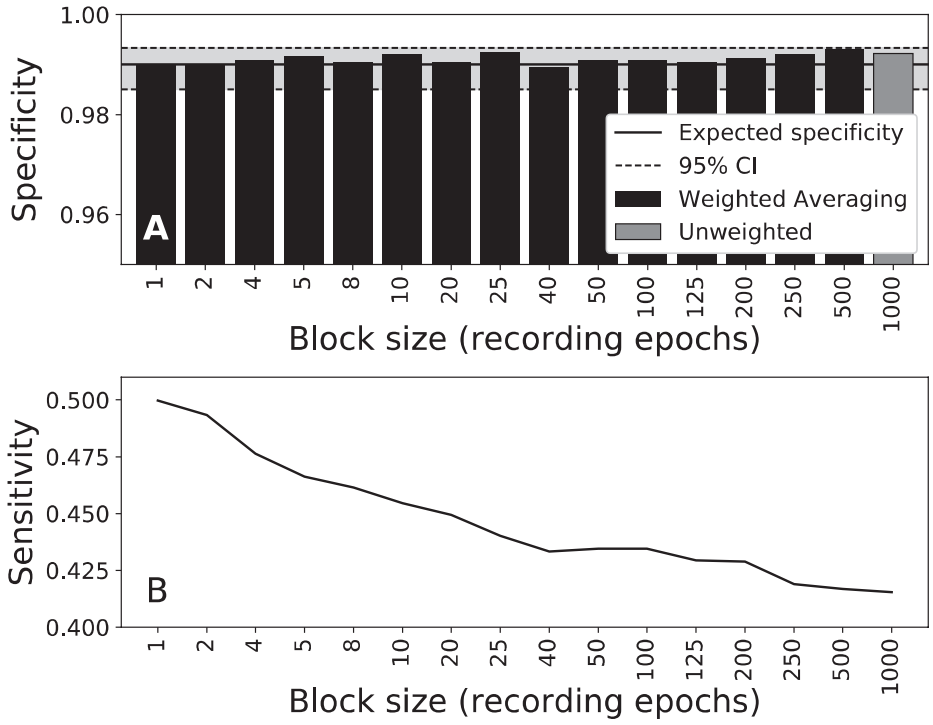

Fig. 6. Using the bootstrap method to control the false positive rate. Fig. 6A shows that the false positive rate, when using the bootstrap method, was within the expected 95% confidence interval across all block sizes, showing clear improvement compared to the equivalent data in Fig. 5 (without bootstrapping). Fig. 6B shows the detection rate for each block size when using the bootstrap statistic with its well-controlled false positive rate (see Fig. 6A). Note that the sensitivity refers to detections across all "response present" data which had a wide range of SNRs.

statistic in the null condition ("response absent") (0.952) was significantly below the expected value of just over one. Following discussion with researchers in the field, it was suggested that the low mean Fmp value may be due to the effects of the analysis window length on the Fmp numerator (Dr Jaime Undurraga, personal communication, 2022). It has already been indicated that the Fmp numerator will not reflect fully the signal power for frequencies below $1/a$ Hz, where a is the length of the analysis window in seconds [3]. For the present study an analysis window length of 14 ms was used, suggesting that the signal power for frequencies below 71.4 Hz would not be fully reflected in the Fmp numerator, biasing the null statistic to be below its expected value. This bias effect has also been reported by the British Society of Audiology [1]. Further simulations were carried out, confirming that the analysis window length can indeed have a bias effect on the numerator of the Fmp, with short analysis window lengths reducing its value (see Figure, Supplemental Digital Content 2). The data shown in the present study must be interpreted in light of the pre-processing techniques used and their associated parameters.

One way to overcome the limitations imposed by the length of the analysis window is to simply increase its length. However, this comes at the expense of potentially reducing the SNR within the coherent average, as most of the ABR signal would be expected to be recorded within the first 15 ms post stimulus [17]. Elberling and Don [3] advocate the selection of a suitable high-pass filter setting to ensure that low-frequency signal components are not excluded from the Fmp numerator as a result of the analysis window length whilst being present in the Fmp denominator. As spectral analyses of the ABR show that most of the signal energy is in the region below 150 Hz [17], caution again must be employed to avoid depreciating the SNR within the coherent average when choosing either an extended analysis window length or higher cut-off frequencies for the high-pass filter.

A further investigation was performed, repeating elements of this study with a raised high-pass filter setting of 100 Hz (see Figures, Supplemental Digital Content 3). In summary, the "VAR Whole Block" method was still more effective than the "VAR MP" method in terms of residual noise reduction and ABR detection (although the difference was less marked – perhaps due to less noise in the EEG limiting the effects of weighted averaging). Inflation of the null Fmp statistic when weighted averaging was applied was also almost fully reduced (apart from for a block size of 1 for the median Fmp level). Weighted averaging will preferentially reduce the frequency components whose amplitudes are greatest in the noise i.e. low frequencies for EEG [6]. This would serve to reduce the bias associated with the finite Fmp analysis window length, minimising 'response absent' Fmp inflation. Fmp inflation for the 'response absent' data was not eliminated fully after the high-pass filter was raised to 100 Hz so additional explanations for its cause should be sought. This finding highlights the importance of avoiding bias in the null Fmp statistic imposed by the finite analysis window length. Supplemental Fig. 3-2 shows the performance of the Fmp in detecting the ABR after weighted averaging. Much higher partial ROC AUC scores were achieved in Supplemental Fig. 3-2 where the high-pass filter was set to 100 Hz, compared with Fig. 2 (main text). This is likely the result of the improved SNR observed by raising the high-pass filter setting. As only one ABR template was used, further research on the combined effects of analysis window length and filter settings on ABR detection is warranted. It was chosen to present the results of the initial 30 Hz high-pass filter setting rather than the better-performing 100 Hz high-pass filter setting, because this filter setting reflects those suggested by the British Society of Audiology guidelines and therefore may better reflect what is happening in current clinical practice when using these recommended filter parameters [1]. The Fmp analysis window parameters recommended by the British Society of Audiology vary by equipment manufacturer and the stimulus used, and are up to 10 ms in duration [1]. This is shorter than the 14 ms analysis window used in the present study, and so the bias present in the Fmp numerator may be larger. This study shows that caution is required in selecting the Fmp analysis window length and filter settings as these interact complexly to affect the Fmp statistic (both when weighted and unweighted) and likely contributed to the inflation in the "response absent" Fmp statistic observed when weighted averaging was applied.

Signal processing methods such as weighted averaging may produce unanticipated effects [18], potentially introducing additional violations

to the assumptions underlying statistical detection methods. Bootstrapping is an effective technique for controlling the false positive rate when the extent of these violations cannot be predicted or controlled [2]; bootstrapping successfully controlled the false positive rate and allowed the benefits of lower block sizes to be harnessed by effectively controlling the false positive rate for all block sizes.

4.2. Optimisation of the noise estimation method

Effective weighting relies upon accurate estimation of the noise levels within each block of recording epochs. Don & Elberling [8] advocated the technique of estimating the noise level by calculating the variance across all samples contained within up to eight equally spaced columns within the block of recording epochs. It is difficult to estimate the number of independent samples in each epoch and so using a large number (e.g. all) sample points to estimate the noise level ensures that no information is neglected. In the current study we compared two methods of estimating the noise levels within a block of recording epochs, the "VAR MP" method and the "VAR Whole Block" method. The "VAR Whole Block" method was able to provide a more accurate estimate of the noise levels within each block of recording epochs, and therefore resulted in lower residual noise levels within the coherent average and better ABR detection (for block sizes of 2 to 10 epochs), compared to the "VAR MP" method. The "VAR MP" method recognises the assumption that the ABR signal is deterministic and calculates the variance for each chosen column of the block of epochs accordingly, before finally averaging together the results. The "VAR Whole Block" method, on the other hand, assumes that the mean value is constant for all values in the block, which allows all samples to be included in a single estimate of variance, making more efficient use of the available data, and increasing the degrees of freedom of the variance estimate. This reduces the random estimation error of the noise variance but at the cost of a bias error (in the "response present" condition) given by the variance of the ABR signal. The magnitude of the impact of this bias is expected to be small for low SNR signals such as the ABR [5]. This perspective would need to be reconsidered if applying the "VAR Whole Block" method to other evoked potential tests which may produce a typically higher SNR [5].

4.3. Optimisation of the block size parameter

Previous work optimising block size for residual noise reduction by Don & Elberling [8] evaluated block sizes of 32, 64, 128, and 256 epochs-per-block and found the smallest block size to be most effective. The current study found even smaller block sizes to be more effective yet. In conjunction with using the "VAR Whole Block" noise estimation method, weighted averaging with a block size of 25 resulted in the largest mean reduction in the residual noise levels in the coherent average (relative to unweighted averaging). Whilst median residual noise levels decreased further for smaller block sizes, mean residual noise levels increased to levels above the baseline level (unweighted averaging). This effect of increased residual noise levels using small block sizes was also observed by Riedel et al. [19]. They found that applying iterative averaging substantially mitigated this increase in residual noise when using small block sizes. Riedel et al. [19] found that applying iterative averaging reduced the optimal block size from 32 to 4 when measuring mean residual noise using simulated data. However, the small number of recordings used limits the ability to meaningfully infer an optimal block size.

In terms of automated ABR detection using the Fmp, the largest partial ROC AUC score was observed using a block size of 2 epochs-per-block (Fig. 2). Whilst this block size was associated with considerable Fmp inflation in "response absent" data when using weighted averaging and a mean *increase* in residual noise levels, overall detection performance was improved (due to an even greater Fmp inflation in the "response present" data). The use of conventional analysis using F-statistics with $\nu_1 = 5$ df led to an excessively conservative response detector, however the false positive rate was well-controlled using the bootstrap method. Despite better overall detection performance, the increase in mean residual noise levels when using small block sizes suggests that a block size of around 25 epochs per block, may be recommended. This provided consistent improvement in ABR quality and also considerable benefit in detection performance, compared to unweighted averaging.

Ultimately, the exact value of the block size parameter will be influenced both by the intended purpose of weighted averaging and the characteristics of the EEG data that the technique is being applied to, including as a result of the recoding parameters selected. The results presented and any derived recommendation relate specifically to the recording parameters used in this study, and may not generalise across other recording parameter configurations. For example, the optimal block size was found to vary depending on the high-pass filter setting used (compare Fig. 1 and Fig. 2 with those presented in Supplemental Digital Content 3). The choice of block size may vary depending on the signal characteristics, including the degree of stationarity and dependence between samples, as well as whether the primary objective is to obtain an averaged waveform with low mean square error for visual inspection or whether weighted averaging is to be combined with a particular statistical detection method. As weighted averaging using small block sizes was associated with a steep drop-off in performance, cautious selection of larger block sizes may be recommended, allowing for differences in EEG characteristics between datasets. Further experimental work with a large sample of ABR data, preferably recorded from the intended target clinical population using the intended recording parameters, should be carried out before finalising recommended protocols for specific applications.

An effective method of controlling for violations to statistical test assumptions is to combine the statistical test with the bootstrap method. This allowed the benefits to ABR detection of using much smaller block sizes to be harnessed. Using the bootstrap to control the false positive rate, the highest detection rate achieved was using one epoch-per-block (50.0%). The lowest detection rate was achieved using unweighted averaging (41.5%) (Fig. 6). This corresponds to just over a 20% relative increase in detection rate. Whilst ABR detection may improve overall using a block size of one epoch-per-block, this parameter value may be undesirable due to the increase in mean residual noise level observed, especially if visual inspection is being used.

4.4. Limitations and future work

The findings in this study are derived from a single (large) database of empirically obtained background EEG data and simulated "response present" data using one ABR template. Further work to evaluate the methods in this study using different databases of EEG data is required in order to determine the generalisability of the presented findings. Future work should also include evaluating the combined performance of weighted averaging with other statistical detection methods such as Hotelling's T^2 test, which previous work has found to be more sensitive than the Fmp [2].

The interactions between data pre-processing techniques and their parameters such as filter settings, weighted averaging, and objective detection methods are complex. Further investigation into the optimal combination of analysis window, filter settings and weighted averaging parameters is therefore warranted, especially prior to clinical implementation of any recommendations.

Future work may also seek to compare the performance of weighted averaging with other published denoising methods such as Kalman-weighted averaging [20], wavelet analysis [21], and adaptive Kalman filtering [22].

5. Conclusion

Weighted averaging provides an effective method for reducing the residual noise level within the averaged waveform. The current work showed that weighted averaging using optimised methods and parameters led to improved ABR detection using the Fmp statistical test on the dataset evaluated, compared to unweighted averaging. The "VAR Whole Block" method for estimating the noise level within each block was able to provide better performance than the "VAR MP" method, especially for smaller block sizes. However, when using smaller block sizes combined with the Fmp statistical detection method, weighted averaging produced an increase in the "response absent" Fmp statistic, relative to unweighted averaging. The Fmp analysis window length, in combination with the high-pass filter setting, introduced a bias to the Fmp statistic. This contributed to the "response absent" Fmp inflation observed. This study highlights the importance of selecting an appropriate Fmp analysis window length and high-pass filter setting, both when calculating the unweighted Fmp statistic, and when combining the Fmp with weighted averaging. Further work is required to optimise the Fmp analysis window length, filter settings, and weighted averaging parameters (block size) in combination, as recommended values for these parameters cannot be considered in isolation. The current work clearly demonstrates the potential of the approach taken, however, further analysis of larger datasets reflecting the specific clinical purpose (e.g. ABR detection in neonates) and the intended recording parameter settings should be carried out before exploiting the benefits the methods presented can provide.

CRediT authorship contribution statement

Richard M. McKearney: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Steven L. Bell:** Supervision, Conceptualization, Writing – review & editing. **Michael A. Chesnaye:** Writing – review & editing. **David M. Simpson:** Supervision, Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The ABR data used in this study are openly available in the University of Southampton Institutional Repository at https://doi.org/10.5258/SOTON/D0168.

Acknowledgements

This work was supported by a studentship from the University of Southampton. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The authors are thankful to Sara M. K. Madsen and James M. Harte for allowing use of the no-stimulus EEG data and to Debbie Cane for collecting the ABR data. Thank you to Dr Jaime Undurraga for the advice regarding the Fmp analysis window length. The authors gratefully acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton. The authors would like to thank the anonymous reviewers for their helpful feedback on the original manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bspc.2023.104676.

References

[1] British Society of Audiology, Recommended Procedure: Auditory Brainstem Response (ABR) testing in Babies, British Society of Audiology, 2019. www.thebsa.org.uk.
[2] M.A. Chesnaye, S.L. Bell, J.M. Harte, D.M. Simpson, Objective measures for detecting the auditory brainstem response: comparisons of specificity, sensitivity and detection time, Int. J. Audiol. 57 (2018) 468–478.
[3] C. Elberling, M. Don, Quality estimation of averaged auditory brainstem responses, Scand. Audiol. 13 (1984) 187–197.
[4] W.H. Martin, J.W. Schwegler, A.L. Gleeson, Y.B. Shi, New techniques of hearing assessment, Otolaryngol. Clin. North Am. 27 (1994) 487–510.
[5] L. Sörnmo, P. Laguna. Bioelectrical Signal Processing in Cardiac and Neurological Applications, Academic Press, Burlington, MA, 2005.
[6] M. Hoke, B. Ross, R. Wickesberg, B. Lütkenhöner, Weighted averaging - theory and application to electric response audiometry, Electroencephalogr. Clin. Neurophysiol. 57 (1984) 484–489.
[7] C. Elberling, O. Wahlgreen, Estimation of auditory brainstem response, abr, by means of bayesian inference, Scand. Audiol. 14 (1985) 89–96.
[8] M. Don, C. Elberling, Evaluating residual background noise in human auditory brain-stem responses, J. Acoust. Soc. Am. 96 (1994) 2746–2757.
[9] S.M.K. Madsen, Accuracy of averaged auditory evoked potential amplitude and latency estimates, [Master's thesis, Technical University of Denmark], 2010.
[10] S.M.K. Madsen, J.M. Harte, C. Elberling, T. Dau, Accuracy of averaged auditory brainstem response amplitude and latency estimates, Int. J. Audiol. 57 (2018) 345–353, https://doi.org/10.1080/14992027.2017.1381770.
[11] J. Lv, D.M. Simpson, S.L. Bell, Objective detection of evoked potentials using a bootstrap technique, Med. Eng. Phys. 29 (2007) 191–198.
[12] D.K. McClish, Analyzing a portion of the ROC curve, Med. Decis. Mak. 9 (1989) 190–195.
[13] A.M. Mood, F.A. Graybill, D.C. Boes, Introduction to the Theory of Statistics, 3rd ed., McGraw-Hill, 1974.
[14] E.B. Wilson, Probable Inference, the Law of Succession, and Statistical Inference, J. Am. Stat. Assoc. 22 (1927) 209–212, https://doi.org/10.1080/01621459.1927.10502953.
[15] M.A. Chesnaye, S.L. Bell, J.M. Harte, L.B. Simonsen, A.S. Visram, M.A. Stone, K. J. Munro, D.M. Simpson, Efficient detection of cortical auditory evoked potentials in adults using bootstrapped methods, Ear Hear. 42 (2020) 574–583.
[16] M.A. Chesnaye, S.L. Bell, J.M. Harte, D.M. Simpson, Controlling test specificity for auditory evoked response detection using a frequency domain bootstrap, J. Neurosci. Methods. 363 (2021), 109352.
[17] J.W. Hall, New handbook of auditory evoked responses, Pearson, Boston, 2007.
[18] B. Lütkenhöner, M. Hoke, C. Pantev, Possibilities and limitations of weighted averaging, Biol. Cybern. 52 (1985) 409–416.
[19] H. Riedel, M. Granzow, B. Kollmeier, Single-sweep-based methods to improve the quality of auditory brain stem responses, Zeitschrift Für Audiol. Audiol. 40 (2001) 82–85.
[20] B. Cone, L.W. Norrix, Measuring the advantage of kalman-weighted averaging for auditory brainstem response hearing evaluation in infants, Am. J. Audiol. 24 (2015) 153–168.
[21] A.P. Bradley, W.J. Wilson, On wavelet analysis of auditory evoked potentials, Clin. Neurophysiol. 115 (2004) 1114–1128.
[22] H. Zhang, M. Zhu, Y. Jiang, D. Wang, X. Wang, Z. Yang, W. Huang, S. Chen, G. Li, A robust extraction approach of auditory brainstem response using adaptive Kalman filtering method, IEEE Trans. Biomed. Eng. 1–1 (2022).