Gaussian Processes for hearing threshold estimation using Auditory Brainstem Responses

Chesnaye M.A., Simpson D.M., Schlittenlacher J. & Bell S.L.

***Abstract*— The Auditory Brainstem Response (ABR) plays an important role in diagnosing and managing hearing loss, but can be challenging and time-consuming to measure. Test times are especially long when multiple ABR measurements are needed, e.g., when estimating hearing threshold at a range of frequencies. While many detection methods have been developed to reduce ABR test times, the majority were designed to detect the ABR at a single stimulus level and do not consider correlations in ABR waveforms across levels. These correlations hold valuable information, and can be exploited for more efficient hearing threshold estimation. This was achieved in the current work using a Gaussian Process (GP), i.e., a Bayesian approach method for non-linear regression. The function to estimate with the GP was the ABR's amplitude across stimulus levels, from which hearing threshold was ultimately inferred. Active learning rules were also designed to automatically adjust the stimulus level and efficiently locate hearing threshold. Simulation results show test time reductions of up to ∼50% for the GP compared to a sequentially applied Hotelling's T^2 test, which does not consider correlations across ABR waveforms. A case study was also included to briefly assess the GP approach in ABR data from an adult volunteer.**

***Index Terms*— Auditory brainstem responses, Gaussian Process, active learning, hearing threshold estimation**

I. Introduction

The Auditory Brainstem Response (ABR) is a brief change in brain activity following the onset of an acoustic stimulus [1]. It is usually measured non-invasively using scalp electrodes (i.e. electroencephalography, or EEG), and comprises a series of peak and trough voltage amplitudes that fall within a short time interval following stimulus onset. It was initially reported back in 1971 [2], and was soon recognised as an effective tool for conducting hearing tests [3] and diagnosing auditory tumours [4]. Nowadays, the ABR plays a central role in evaluating hearing in newborns and others who may not be able to provide behavioural responses [5] and can be used to diagnose a wider range of neurological disorders [6].

The focus for the current work is on ABR hearing threshold estimation, where ABR hearing threshold is defined as the lowest stimulus level that evokes an ABR. The main target group is newborns with suspected hearing loss, i.e. those who failed the initial screening test. In the United Kingdom, this includes approximately 16,000 newborns per annum [7]. The overarching aim for ABR hearing threshold estimation is to specify hearing loss characteristics and facilitate the subsequent management of hearing loss, e.g. by fitting a hearing aid.

Due to time constraints in the clinic, ABR hearing thresholds are usually estimated for just the frequencies deemed essential for speech comprehension, which include 500, 1000 2000 and 4000 Hz [8]. Even when estimating just 8 frequencies (4 per ear), average test times were previously estimated to be in the 30 to 60 minute range for sleeping newborns [9,10]. For newborns who were restless, test times were slightly longer, and testing occasionally had to be stopped, resulting in incomplete information for the subsequent management of hearing loss [10]. More efficient methods for ABR hearing threshold estimation are thus desirable, as this would allow more information to be obtained within the available test time, ultimately facilitating diagnosis and treatment of hearing loss.

Long test times for ABR hearing threshold estimation are primarily due to poor signal-to-noise ratios (SNRs) of the ABR. With a typical peak-to-trough amplitude (PTTa) of around 1 μV or less (e.g. [11, 12]), the ABR tends to be hidden in the much larger EEG background activity, which can be in the ∼10 μV range or more, even after pre-processing and artefact rejection [13]. In order to detect the ABR reliably, it's SNR therefore first needs to be improved, achieved by presenting many stimuli to the subject (up to several thousand) and averaging across the short time intervals following stimulus onset. An experienced clinician is then given the task to visually inspect the averaged waveform to determine whether an ABR is present or absent [5].

While visual inspection can potentially be quite sensitive, studies have shown that results can vary both within and between examiners [14, 15]. This might be due to differences in expertise and equipment, or variations in concentration levels and fatigue. Either way, findings suggest that visual inspection introduces a subjective, error-prone element to the analysis, potentially compromising the accuracy and efficiency of the test. Accordingly, researchers have focussed on automating ABR detection procedures, leading to a plethora of statistical approaches and machine learning techniques for response classification and/or determining the presence or absence of ABRs [e.g., 16-45, just to name a few].

With respect to objective methods for ABR hearing threshold estimation in the clinic, these typically comprise three main components: (1) a statistical test or machine learning technique for determining the presence or absence of an ABR, (2) a sequential test strategy for specifying when and how often to analyse data whilst also rigorously controlling error

rates, especially false positives (i.e., detection of ABR when none are present), and (3) a protocol for adjusting the stimulus level and homing in on hearing threshold. The vast majority of research has focussed on (1), i.e., the test statistic and/or machine learning technique [e.g. 16-38], which is arguably also the most impactful regarding test accuracy and efficiency. A limitation, however, is that most methods implicitly assume independence between stimulus levels, meaning information from previously measured waveforms is discarded. This likely leads to a sub-optimal test performance, as ABRs are known to be correlated across stimulus levels and frequencies [1]. Although some researchers have leveraged these correlations through "curve-fitting" strategies [26-28] (see also the Discussion), these approaches are typically applied as a post-hoc analysis (i.e., after all data has been collected), and thus disregard the sequential testing aspect involved in real-time hearing threshold estimation in the clinic.

With respect to (2), i.e., sequential testing, this is important for providing prompt feedback to clinicians, and helps to keep test time low as data collection can be stopped as soon as an ABR is deemed present or absent. The challenge is that repeated hypothesis testing increases the probability of making an error, known as an inflated false-positive rate (FPR) [46]. In order to prevent inflated FPRs, and preserve the significance level of the test, the critical thresholds for response detection need to be adjusted, for which various methods have been proposed [39-44]. However, in order to find these critical thresholds, limitations need to be imposed on how long and/or how often data can be analysed, which is problematic for ABR detection due to the ABR's unknown SNR. In particular, any pre-determined sample size will tend to result in either an over- or an under-powered test, leading to prolonged test times and reduced test sensitivities, respectively.

Finally, with respect to (3), i.e., the stimulus selection protocol, this is needed to efficiently switch between levels, and quickly home in on hearing threshold. The most common approach currently used is an "X-down-Y-up" strategy, where the X and Y refer to the dB change in stimulus level following a detection and a non-detection, respectively [45]. A potential drawback for this approach is its susceptibility to false-positives, which can have a large impact on test accuracy (see also the Discussion).

The main goal for the current work was to develop an automated approach for objective ABR hearing threshold estimation, and to overcome some key limitations underlying existing methods. The approach revolves around the Gaussian Process (GP), which is a Bayesian approach for non-linear regression [47]. The function to estimate by the GP was furthermore the ABR's amplitude-intensity growth function, i.e., ABR amplitude as a function of the stimulus level. Active learning rules were also designed to automatically switch between stimulus levels and quickly home in on hearing threshold.

The GP is attractive, firstly because it provides a rigorous framework for learning and exploiting the correlation structure underlying the ABR data. Secondly, as a Bayesian approach that does not utilize repeated Frequentist hypothesis testing, advanced sequential test strategies for preventing inflated FPRs are no longer needed, which greatly simplifies the sequential analysis and provides much needed flexibility in terms of how long and how often data can be analysed. Thirdly, the GP provides a single framework for evaluating data from multiple stimulus levels, which helps to mitigate the impact of any single false-positive or data outlier (see also the Discussion). Finally, the GP is not confined to sequential "up-down" stimulus selection protocols, and novel active learning rules for automatically adjusting the stimulus level may offer new opportunities for improving the efficiency and accuracy of the test.

In what follows, the GP with application to ABR hearing threshold estimation is described in more detail (Sections II and III). A sequentially applied Hotelling's T^2 (HT^2) test is also presented (Section IV), which does not consider correlations across ABR waveforms. The FPR for the sequential HT^2 test was controlled, per stimulus level, using a sequential test strategy from [42,43], called the Convolutional Group Sequential Test (CGST; see also Section IV). The CGST previously demonstrated an efficient test performance for ABR detection [42], and thus provides a challenging benchmark to compare against. The sequential HT^2 test was also combined with a 10-down-10-up test strategy (i.e., the stimulus was adjusted in 10 dB steps) for automatically homing in on hearing threshold. The performance of the GP and the sequential HT^2 test was evaluated extensively in simulated data, which emulates a wide range of sensorineural and conductive hearing loss configurations (Section V). Simulated data were chosen in this study as this allows large, well-controlled data sets to be generated, providing a powerful assessment of test performance for a wide range of test conditions. As a proof of concept, the GP approach was also briefly evaluated in ABR data recorded from an adult volunteer (Section VI). Finally, various GP test parameters are considered in the Discussion, along with alternative methods from the literature, study limitations, and directions for future work.

II. Gaussian Processes

This section begins with an informal description of the Gaussian Process (GP) supported by illustrations provided in Fig. 1. The informal description is aimed at readers without expertise in the field, and aims to build an intuitive understanding of the GP. Following the informal description is a brief overview of the GP's mathematical framework.

A. Understanding Gaussian Processes

The GP can be viewed as a model of our beliefs, and more specifically, a model of our beliefs regarding some function of interest, denoted by $f(x)$. In the context of ABR hearing threshold estimation, $f(x)$ refers to the ABR wave V PTTa growth function, defined as the voltage difference between the ABR wave V peak and trough (e.g. [48]) as a function of stimulus level x. It is worth emphasizing here that the GP aims to model the true (i.e., noise-free) PTTa function values, not the observed (noisy) PTTa values measured from data.

The GP models our beliefs of $f(x)$ using a multivariate normal (MVN) distribution, which comprises just two components: (1) a vector of means, and (2) a covariance matrix. The

TABLE I
A NOTATION TABLE FOR THE MOST IMPORTANT SYMBOLS USED THROUGHOUT THIS WORK.

Symbol	Description
$f(x)$	function to be estimated by the GP
x_n	the n^{th} stimulus level to test at
$\boldsymbol{X}_P$	vector containing P locations (stimulus levels) along which $f(x)$ is estimated
$\boldsymbol{X}_T$	vector containing T stimulus levels to test at
$\boldsymbol{O}_T$	vector containing T observed PTTa values
$\boldsymbol{\sigma}^2_T$	vector containing the T variances associated with $\boldsymbol{O}_T$
$\boldsymbol{\mu}_P$	prior mean PTTa values along $\boldsymbol{X}_P$
$\boldsymbol{\mu}_T$	prior mean PTTa values along $\boldsymbol{X}_T$
$\boldsymbol{\Sigma}_P$	prior covariance matrix for PTTa values along $\boldsymbol{X}_P$
$\boldsymbol{\Sigma}_T$	prior covariance matrix for PTTa values along $\boldsymbol{X}_T$
$\boldsymbol{\Sigma}_{PT}$	prior covariance matrix for PTTa values along $\boldsymbol{X}_P$ and $\boldsymbol{X}_T$
$\boldsymbol{\Sigma}_{TP}$	prior covariance matrix for PTTa values along $\boldsymbol{X}_T$ and $\boldsymbol{X}_P$
$\bar{\boldsymbol{\mu}}_P$	posterior mean PTTa values along $\boldsymbol{X}_P$
$\bar{\boldsymbol{\Sigma}}_P$	posterior covariance matrix for PTTa values along $\boldsymbol{X}_P$
$\boldsymbol{D}_n$	the coherently averaged epoch at the n^{th} stimulus level
G_n	the number of recorded epochs at the n^{th} stimulus level
a_n	biased estimate of the PTTa value at the n^{th} stimulus level
o_n	unbiased estimate of the PTTa value at the n^{th} stimulus level
T_i	the i^{th} PTTa target for the GP to locate
x_{T_i}	the estimated stimulus level for evoking T_i
δ_i	the required level of certainty (the standard deviation of the GP posterior) before x_{T_i} is deemed located

vector of means represents our belief regarding the most likely PTTa value at each stimulus level, and the covariance matrix captures the level of uncertainty surrounding the mean values. The covariance matrix also encodes expectations of "function smoothness", i.e., the rate at which PTTa values are expected to change with x (further clarified below).

To give an example, consider **Fig. 1, panel (a)**, which shows a simplified depiction of a GP prior. The GP prior describes the space within which $f(x)$ is expected to reside before having observed data. The thick dashed line in panel (a) represents the vector of means, which was set to zero for all stimulus levels, essentially representing the assumption that subject is deaf. The shaded region then represents uncertainty surrounding these mean values, and was defined by ± 2.575 standard deviations from the mean, representing the mean's 99% confidence intervals. Note that the standard deviations are given by the main diagonal of the covariance matrix. The $f(x)$ function values that the GP aims to estimate are also shown in panel (a) as a gray dotted line. Note that the GP prior in Fig. 1, panel (a) covers a relatively large space, representing our initial uncertainty regarding the true $f(x)$ function values.

As data becomes available, our understanding of $f(x)$ evolves, which is taken into account by transforming the GP prior into a GP posterior. This process considers the observed data, but also prior assumptions regarding function smoothness. To clarify, consider **Fig. 1, panel (b)** where a PTTa of $\sim$0.62 μV was observed at 70 dB Hearing Level (HL, albeit simulated). How this data impacts on the GP posterior depends firstly on how reliable (i.e. how noisy) it is. In this case, the $\sim$0.62 μV PTTa value was relatively noisy, which implies that there is still uncertainty regarding the true (noise-free) $f(70)$ value. The 99% confidence intervals for the GP posterior at $f(70)$ in panel (b) are therefore still relatively wide. How the $\sim$0.62 μV observation impacts on the posterior also depends on the assumed covariance structure. The latter is represented by the off-diagonal elements of the covariance matrix, and leads to information being "smeared" across stimulus levels. The observation at 70 dB HL, for example, impacted not only on our expectations for $f(70)$, but also on our expectations for the adjacent stimulus levels.

The assumed covariance structure is an important component of the GP, and provides the foundation for the GP's efficacy at conducting non-linear regression. If the covariance structure is assumed in advance, it should therefore be chosen carefully. Alternatively, the covariance structure can be learned from the data, which is further considered in Section III.

Finally, **panels (c)** and **(d)** in **Fig. 1** show two additional GP Posteriors. These panels aim to illustrate the active learning rules (how to choose the next stimulus level) and stopping criteria, and are further considered in Section III under "Illustrative example". To briefly explain here also: The active learning rules first aim to obtain a roughly monotonic estimate of $f(x)$ (Fig. 1, panel c), after which the focus shifts to reducing uncertainty along the $0.1 \leq f(x) \leq 0.5$ μV interval, but prioritising the lower amplitude region. Fig. 1, panel (d) shows the final GP posterior from which hearing threshold was inferred. Further details are presented in Section III.

B. Mathematical framework of Gaussian Processes

This section defines the mathematical framework underlying the GP [47, 49]. In theory, the GP provides a means to estimate $f(x)$ for all conceivable values of x, resulting in an infinite set of function values for the GP to estimate. In practice, the GP is of course confined to estimating $f(x)$ at just a finite set of locations. These locations will henceforth be referred to as the "prediction locations", and are denoted by $\boldsymbol{X}_P$ with elements x_j for j=1, 2, ..., P.

As mentioned previously, the GP starts with an initial MVN distribution called the prior, which is fully defined by just two components: a P-dimensional vector of means, denoted by $\boldsymbol{\mu}_P$, and a P x P dimensional covariance matrix, denoted by $\boldsymbol{\Sigma}_P$, defined as:

$$\boldsymbol{\mu}_P = \mathbb{E}[f(\boldsymbol{X}_P)] \tag{1}$$

$$\boldsymbol{\Sigma}_P = \mathbb{E}[(f(\boldsymbol{X}_P) - \boldsymbol{\mu}_P)(f(\boldsymbol{X}_P) - \boldsymbol{\mu}_P)'] \tag{2}$$

where $'$ denotes transpose. The $\boldsymbol{\mu}_P$ and $\boldsymbol{\Sigma}_P$ components can be specified through a "mean function" and a "covariance function", respectively, and are further considered in Section III.

After having defined the prior, data is collected by probing $f(x)$ at a set of T locations. These locations are referred to as the "test locations", and are denoted by $\boldsymbol{X}_T$ with elements x_n for n = 1, 2, ..., T. Note that $\boldsymbol{X}_T$ may differ from $\boldsymbol{X}_P$. Probing $f(x)$ at $\boldsymbol{X}_T$ gives a T-dimensional column vector of observations, say $\boldsymbol{O}_T$ with elements o_n for n = 1, 2, ..., T. The GP then uses prior assumptions and information from $\boldsymbol{O}_T$ to transform the MVN prior into a MVN posterior, thus reshaping the space within which $f(\boldsymbol{X}_P)$ is expected to reside.

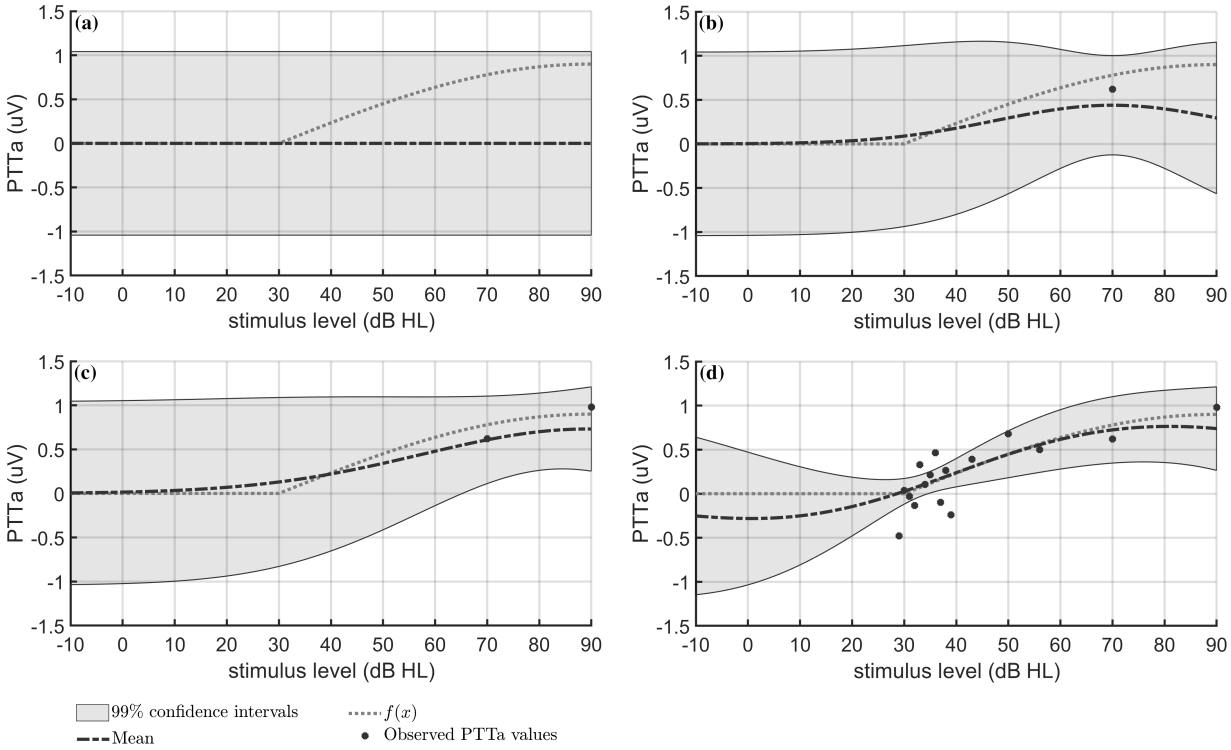

Fig. 1. An illustration of the Gaussian Process (GP) for peak-to-trough amplitude (PTTa) growth function estimation. The GP is characterized by a multivariate normal (MVN) distribution, which defines the space in which the growth function, denoted by $\boldsymbol{f}(\boldsymbol{x})$, is expected to reside. Before having observed data, this space may be vast, reflecting our initial uncertainty about the true $\boldsymbol{f}(\boldsymbol{x})$ function values. As data is collected and uncertainty regarding $\boldsymbol{f}(\boldsymbol{x})$ is reduced, the MVN space is transformed, reflecting our improved understanding of $\boldsymbol{f}(\boldsymbol{x})$. This transformation considers prior assumptions as well as the observed data. **Panel (a)**: A simplified depiction of a GP prior, i.e., the MVN space before having observed data. The thick dashed line shows the MVN mean and the shaded regions are defined by the 99% confidence intervals of the (marginal) MVN distribution. The $\boldsymbol{f}(\boldsymbol{x})$ function, to be estimated by the GP, is also shown as a dotted line. **Panel (b)**: A simplified depiction of a GP posterior, i.e., the MVN space after observing a PTTa value of 0.62 μV at 70 dB HL. **Panel (c):** The GP posterior after also having observed a PTTa of 0.98 μV at 90 dB HL, and **Panel (d):** The GP posterior after all stopping criteria were met. Details regarding the stopping criteria, the prior assumptions, and the active learning rules for stimulus selection are presented in Section III.

It is worth noting that the MVN posterior is defined across the same $\boldsymbol{X}_P$ locations as the MVN prior.

In order to actually derive the posterior, the following additional components should be defined: (1) the T-dimensional column mean vector for the test locations, denoted by $\boldsymbol{\mu}_T$, (2) the T x T dimensional covariance matrix for the test locations, denoted by $\boldsymbol{\Sigma}_T$, (3) the T x P dimensional cross-covariance matrix for the test locations and the predictions locations, denoted by $\boldsymbol{\Sigma}_{TP}$, and similarly (4), the P x T dimensional cross-covariance matrix for the prediction locations and the test locations, denoted by $\boldsymbol{\Sigma}_{PT}$. These components are all specified through the aforementioned mean and covariance functions, and are further considered in Section III.

If the $\boldsymbol{O}_T$ observations are noisy, then the last step before deriving the posterior is to add an additional component to the diagonal of $\boldsymbol{\Sigma}_T$. In particular, the main diagonal of $\boldsymbol{\Sigma}_T$ is respecified as $diag(\boldsymbol{\Sigma}_T) + \boldsymbol{\sigma}_T^2$ where $diag$ denotes the main matrix diagonal, and $\boldsymbol{\sigma}_T^2$ is a T-dimensional vector containing the estimated variances associated with the $\boldsymbol{O}_T$ measurements [47].

Finally, the equations for generating the MVN posterior are given by [47]:

$$\bar{\boldsymbol{\mu}}_P = \boldsymbol{\mu}_P + \boldsymbol{\Sigma}_{PT}\boldsymbol{\Sigma}_T^{-1}(\boldsymbol{O}_T - \boldsymbol{\mu}_T) \tag{3}$$

and

$$\bar{\boldsymbol{\Sigma}}_P = \boldsymbol{\Sigma}_P - \boldsymbol{\Sigma}_{TP}\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Sigma}_{PT} \tag{4}$$

where $\boldsymbol{\Sigma}_T^{-1}$ is the inverse of $\boldsymbol{\Sigma}_T$. These equations are the result of deducing a conditional MVN distribution from a joint MVN distribution [49]. Note that $\bar{\boldsymbol{\mu}}_P$ and $\bar{\boldsymbol{\Sigma}}_P$ are repeatedly recalculated as new data arrives, which involves adding entries to $\boldsymbol{O}_T$ and $\boldsymbol{\sigma}_T^2$, and recomputing $\boldsymbol{\Sigma}_T$, $\boldsymbol{\Sigma}_{TP}$ and $\boldsymbol{\Sigma}_{PT}$.

III. ABR HEARING THRESHOLD ESTIMATION USING GAUSSIAN PROCESSES

The overarching aim for the GP in the current work was to infer hearing threshold from the GP-estimated PTTa growth function. The following sections provide a more detailed description of the PTTa (Section III.A) along with the mean and covariance functions for specifying the priors (Section III.B). The active learning rules for adjusting the stimulus level are then also presented (Section III.C) along with the criteria for deciding when to stop data collection and infer hearing threshold (Section III.C).

A. Peak-to-trough amplitude estimation

This section describes the approach for estimating the PTTa values, which comprise the elements of $\boldsymbol{O}_T$, i.e., the o_n values for $n = 1, 2, ..., T$. As the o_n estimates are noisy, they need to be provided with an estimate of variance. These variances comprise the elements of $\boldsymbol{\sigma}_T^2$, i.e., the σ_n^2 values for $n = 1, 2, ..., T$.

As mentioned previously, the ABR has a poor SNR, and many waveforms need to be averaged before an ABR can reliably be detected. These waveforms are referred to as epochs, and are given by the brief EEG intervals following the stimuli. At each stimulus level, the recorded epochs can be considered as a matrix:

$$\boldsymbol{D}_n = \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,L} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ d_{G_n,1} & d_{G_n,2} & \cdots & d_{G_n,L} \end{pmatrix}$$

where $d_{i,j}$ denotes the j^{th} sample of the i^{th} epoch, L is the number of EEG samples within each epoch, and G_n is the number of epochs (the ensemble size) recorded at the n^{th} stimulus level. Averaging down each of the L columns of $\boldsymbol{D}_n$ gives the coherently averaged epoch, say $\bar{\boldsymbol{D}}_n$, from which a PTTa can be estimated.

Estimating the PTTa requires identifying wave V peak and trough, and then computing the difference [48]. The challenge is that peak and trough latencies (i.e., time following stimulus onset) vary depending on factors such as the stimulus level, measurement technique, and filtering parameters. Thus, to ensure that the peak and trough can be located, a relatively wide search window is needed. However, employing a wide search window increases the likelihood of detecting artefactual maxima and minima, which introduces noise to the estimate and reduces the SNR. In short, it is desirable to keep the search window as short as possible, but to still provide sufficient coverage to ensure robust peak and trough detection. This was facilitated by a sliding window approach.

Sliding window approach

The sliding window approach begins by defining two adjacent 4 ms windows: One for locating the peak and one for locating the trough. These windows are then slid across $\bar{\boldsymbol{D}}_n$ in steps of 2 ms. At each window location, a peak and trough are computed using:

$$P_i = \max_{j \in [\boldsymbol{A_i}]} \bar{\boldsymbol{D}}_n(j) \tag{5}$$

and

$$T_i = \min_{j \in [\boldsymbol{B_i}]} \bar{\boldsymbol{D}}_n(j) \tag{6}$$

where $\boldsymbol{A_i}$ and $\boldsymbol{B_i}$ contain the indices corresponding to the windows for locating the peak and trough, respectively, at step i. For each window location, a peak-to-trough difference is computed, and the final PTTa value, say a_n, is given by the largest of these differences:

$$a_n = \max_{i \in [\boldsymbol{I}]} (P_i - T_i) \tag{7}$$

where $\boldsymbol{I}$ contains the indices for the sliding window positions. A total of five sliding window positions were evalauted ($\boldsymbol{I} = [1, 2, 3, 4, 5]$), with the starting positions of the first window set at [2, 4, 6, 8, 10] ms, respectively (see **Fig. 2** for an exanple). It is worth noting that this approach assumes the peak precedes the trough, and that the interval between peak and trough is less than 8 ms.

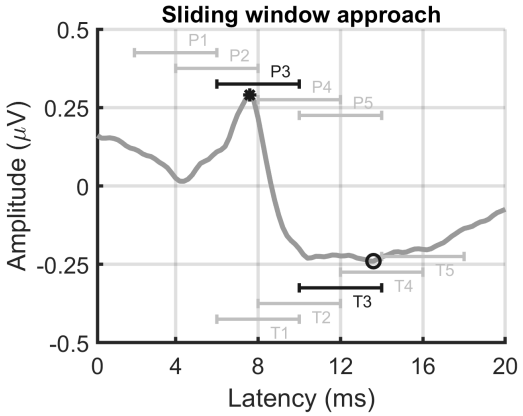

Fig. 2. An illustration of the sliding window approach for peak and trough detection. The approach involves defining two adjacent 4 ms windows - one for locating the peak and another for locating the trough - which are slid across $\bar{\boldsymbol{D}}_{\boldsymbol{n}}$ in steps of 2 ms. At each step, a peak and trough are estimated using Eq. (5) and (6), and a peak-to-trough difference is computed. The final peak-to-trough amplitude equals the largest of these differences and is found using Eq. (7). For the current example, the largest peak-to-trough difference was observed for the 3rd window location: The $\boldsymbol{A_3}$ indices for locating the peak span the 6-10 ms interval, and the largest value along this interval was $\boldsymbol{P_3} = \boldsymbol{0.26\mu}$V, which is indicated in the figure by an asterisk. The $\boldsymbol{B_3}$ indices for locating the trough span the 10-14 ms interval, and the smallest amplitude along this interval was $\boldsymbol{T_3} = \boldsymbol{-0.24\mu}$V, which is indicated in the figure by a circle. The estimated peak-to-trough amplitude in this example was 0.53 $\boldsymbol{\mu}$V.

The PTTa bias

The GP assumes that the $\boldsymbol{O}_T$ measurements are potentially noisy, but unbiased observations of $f(x_n)$. However, this assumption does not hold due to the residual background activity in $\bar{\boldsymbol{D}}_n$, which biases a_n towards over-estimated PTTa values. In particular, the expected value for a_n is given by:

$$\mathbb{E}[a_n] = f(x_n) + \mathbb{E}[b_n] \tag{8}$$

where $\mathbb{E}[b_n]$ represents the bias at stimulus level x_n. It should be stressed that although $\mathbb{E}[b_n]$ is introduced by the residual background activity in $\bar{\boldsymbol{D}}_n$, it is dependent on the SNR of $\bar{\boldsymbol{D}}_n$, which is essentially due to the search process that is involved in peak and trough detection. Since the SNR varies throughout the test, the $\mathbb{E}[a_n]$ estimates also vary, which adversely affects the regression analysis conducted by the GP.

One way to obtain an unbiased estimate of $f(x_n)$ might be to estimate $\mathbb{E}[b_n]$ and subtract this from a_n. However, this is challenging, as the SNR is unknown. This incentivized

a maximum likelihood approach for unbiased PTTa estimation.

Unbiased PTTa estimation

The maximum likelihood approach aims to replace the biased a_n estimates with unbiased estimates, denoted by o_n. The o_n values were estimated using:

$$o_n = \arg \max_{R \in [\boldsymbol{R}]} \phi(a_n \mid R) \tag{9}$$

where $\phi(a_n \mid R)$ is the distribution underlying a_n, evaluated at location a_n, and under the assumption that data contained an ABR with a true (i.e., noise-free) PTTa equal to R. The $\boldsymbol{R}$ vector furthermore contains all noise-free PTTa values to evaluate. In short, the (biased) a_n estimate is replaced with the most likely (unbiased) value in $\boldsymbol{R}$. The main challenge for this approach is to find the $\phi(a_n \mid R)$ distributions, which were approximated using a bootstrap.

The bootstrap

The bootstrap is a random resampling with replacement procedure, which was previously adapted in [18] to approximate null distributions for ABR test statistics. In the current work, however, the aim was to approximate not just the null distribution, but all $\phi(a \mid R)$ distributions (for all $R \in [\boldsymbol{R}]$) where a denotes the axis along which the distribution is defined.

Starting with the null distribution, this was approximated by randomly resampling (with replacement) blocks of EEG measurements from the original recording with no regards to where the stimuli occur [18]. Averaging across the resampled blocks then gives an "incoherent average", as time-locking between the resampled blocks and the stimulus triggers has been disrupted. It is assumed that any residual power of the ABR in the incoherent average is negligible, and hence that the incoherent average represents data under the null hypothesis, representing "no ABR present". Repeating the procedure many times, and calculating a PTTa value from each incoherent average, then gives a distribution of PTTa values, which is assumed to be an approximation of the null distribution.

Approximating the alternative distributions requires a minor tweak, which consists of adding an ABR template waveform (presented in Section V) with a PTTa of R to the incoherent averages, prior to estimating the noisy PTTa values.

With respect to the number of bootstrap repetitions, it is important to strike a balance between achieving sufficiently accurate approximations, and managing computational burden. In the current work, the number of repetitions was set to 2000, as this previously led to accurate results when evaluating test specificity for ABR detection [22] and was sufficiently fast.

The Gaussianity assumption

An important assumption underlying the GP is that the o_n estimates are Gaussian-distributed, which suggests that the o_n values can potentially be negative. How to simulate negative PTTa values, however, is not clear, which implies that the bootstrapped distributions cannot be generated along the $R < 0$ interval. As a result, the o_n estimates also cannot be negative, which violates the Gaussianity assumption. Non-negative o_n estimates bias the GP towards over-estimated $f(x)$ function values, which complicates the hearing threshold estimation procedure.

In order to prevent Gaussianity violations, the bootstrapped distributions along the $R < 0$ interval were extrapolated from those along the $R > 0$ interval. In particular, it was assumed that the distributions along $R < 0$ were shifted versions of those along $R > 0$. More formally:

$$\phi(a \mid -R) = \phi(a + 2\Delta_R \mid R) \tag{10}$$

where Δ_R is the difference between (1) the median value of $\phi(a \mid 0)$, i.e., the bootstrapped distribution under H_0, and (2) the median value of $\phi(a \mid R)$, i.e., the bootstrapped distribution when data contained an ABR with a PTTa equal to R. The distributions along $R < 0$ were hence assumed to be equivalent to those along $R > 0$, just shifted by $2\Delta_R$.

Measurement error

Finally, the GP requires each o_n estimate to be provided with a variance, previously denoted by σ_n^2. If the distribution underlying o_n is known, then estimating σ_n^2 is relatively straightforward. Fortunately, this distribution was already approximated when estimating o_n in Eq. (9), and is given by:

$$\Phi_n(R) = \phi(a_n \mid R) \tag{11}$$

where R can once again assume values as defined in $\boldsymbol{R}$. After approximating $\Phi_n(R)$, σ_n^2 is estimated using:

$$\sigma_n^2 = \left[\frac{Q_n(0.8413) - Q_n(0.1587)}{2} \right]^2 \tag{12}$$

where $Q_n(0.1587)$ and $Q_n(0.8413)$ denote the 0.1587 and 0.8413 quantiles, respectively, of $\Phi_n(R)$. These quantiles correspond to ± 1 standard deviations from the mean of a Guassian distribution.

B. The mean and covariance functions

As mentioned previously, the GP starts with an initial MVN space, called the GP prior, which requires a mean vector and a covariance matrix to be defined. Starting with the mean vector, this was defined under the assumption of no "ABR present", and was therefore set to zero for all stimulus levels, giving $\boldsymbol{\mu}_P = 0$ and $\boldsymbol{\mu}_T = 0$. This essentially represents the belief that subject is deaf, which might be considered a clinically conservative starting position, i.e., it may be safer to err on the side of caution and assume that a careful diagnosis of hearing loss is required, rather than assume subject has normal hearing. The main incentive for assuming a zero mean prior, however, was to facilitate monotonic estimates of $f(x)$, which help to reduce ambiguity in hearing threshold location (further clarified in Section III.C).

With respect to the covariance matrix, this is typically defined using a function, also known as a covariance kernel. In the current work, it was assumed that $f(x)$ function values for adjacent stimulus levels were similar, and that similarity decreased as the distance between stimulus levels increased.

More specifically, covariance was assumed to decay exponentially with the distance in x, which can be modelled using the widely-used exponential covariance kernel [49]. Accordingly, the $\boldsymbol{\Sigma}_P$, $\boldsymbol{\Sigma}_T$, $\boldsymbol{\Sigma}_{T,P}$, and $\boldsymbol{\Sigma}_{P,T}$ covariance matrices were defined as:

$$\boldsymbol{\Sigma}_P = s \cdot \exp^{-\frac{(\boldsymbol{x}_P - \boldsymbol{x}'_P)^2}{\theta}} \tag{13}$$

$$\boldsymbol{\Sigma}_T = s \cdot \exp^{-\frac{(\boldsymbol{x}_T - \boldsymbol{x}'_T)^2}{\theta}} \tag{14}$$

$$\boldsymbol{\Sigma}_{P,T} = s \cdot \exp^{-\frac{(\boldsymbol{x}_P - \boldsymbol{x}'_T)^2}{\theta}} \tag{15}$$

$$\boldsymbol{\Sigma}_{T,P} = s \cdot \exp^{-\frac{(\boldsymbol{x}_T - \boldsymbol{x}'_P)^2}{\theta}} \tag{16}$$

where s is the "scale parameter" and θ is the "length scale parameter".

Starting with the scale parameter, this defines the variance of the MVN distribution, and can be used to specify the range within which the $f(x)$ function values are expected to lie before having observed data. In the current work, it was assumed that 99.9% of the $f(x)$ values were less than 1.25 μV, which was roughly motivated by results from [11, 12]. Two-sided 99.9% confidence intervals coincide with $\pm$3.09 standard deviations from the mean, giving s = $(\frac{1.25}{3.09})^2$ = 0.1636 μV.

With respect to the length scale parameter θ, this controls the rate at which covariance between PTTa values decays as the distance in x increases, and is related to "function smoothness", i.e. how quickly the $f(x)$ function values are expected to change with x. As each subject is expected to have his/her own PTTa growth function, the optimal length scale is expected to vary between individuals. Therefore, instead of assuming θ in advance, θ was treated as a random variable and was estimated from the data using standard maximum likelihood estimation [49]. Pilot simulations (specifics not presented) suggest a favorable test performance when confining θ to the [1000, 2000] interval. A uniform prior for θ was therefore defined on the [1000, 2000] interval.

C. Active learning rules, stopping criteria, and hearing threshold estimation

The overarching aim for the GP is to estimate hearing threshold, which is assumed to be located at the largest x value where $f(x) = 0$, henceforth denoted by x_{HT}. The challenge in locating x_{HT} is that $f(x)$ is zero not just at a single location, but at all inaudible stimulus levels. Consequently, there is a risk that the GP converges on an $f(x) = 0$ location that is below x_{HT}, leading to an under-estimated hearing threshold.

One way to mitigate the risk of converging on false hearing threshold estimates is to approach $f(x_{HT})$ from above, i.e., from the $f(x) > 0$ interval. To facilitate this, active learning rules were designed to first arrive at a monotonic estimate of $f(x)$, which then provides directional guidance on the $f(x) = 0$ location. This was achieved by first locating several non-zero PTTa targets, i.e., locating the x_{T_i} values where $f(x_{T_i}) = T_i$, where T_i denotes the PTTa targets for $i = 1, 2, ..., I$. The x_{T_i} values are furthermore estimated one at a time, starting with the largest T_i target, and only moving on to the smaller targets after locating the larger ones.

The aim for the active learning rules was thus to automatically adjust the next stimulus level, and efficiently locate the x_{T_i} values. The stimulus level was adjusted every 500 epochs (approximately every 10 seconds), and was set to the most likely x where $f(x) = T_i$, albeit under the condition that T_i had not yet been located. More specifically, the next level to test at was given by:

$$x_n = \arg\max_{x_j \in \boldsymbol{X}} \boldsymbol{N}(T_i, \bar{\mu}_{x_j}, \bar{\sigma}^2_{x_j}) \tag{17}$$

where $\boldsymbol{X}$ denotes all potential test locations, and $\boldsymbol{N}(T_i, \bar{\mu}_{x_j}, \bar{\sigma}^2_{x_j})$ is a univariate GP posterior, defined at a single prediction location, given by x_j. Note that this univariate GP posterior is derived using Eq. (3) and (4). Note again that the GP was only allowed to test at x_n if T_i was not yet located. The T_i target was deemed located when the standard deviation of the GP posterior at location x_n was less than δ_i, i.e., for $\bar{\sigma}^2_n < \delta_i$ where δ_i is specified by the user. If this condition was met, then the GP was allowed to estimate the next target, and x_n was instead found for target T_{i-1}. After locating all T_i targets, hearing threshold was estimated using Eq. (17) with T_i set to zero.

1) A caveat: profound hearing loss: In the case of profound hearing loss, it is conceivable that all $f(x)$ function values are smaller than the largest T_i target. In this case, the GP may waste time attempting to locate a target that does not exist. To prevent this, the GP posterior was first inspected at the maximum permitted stimulus level, say x_{max}, and if the most likely PTTa value at x_{max} was smaller than the current T_i target, then x_n was set to x_{max}. The most likely PTTa value at x_{max}, say R_{max}, was found using:

$$R_{max} = \arg\max_{R \in [\boldsymbol{R}]} \boldsymbol{N}(R, \bar{\mu}_{x_{max}}, \bar{\sigma}^2_{x_{max}}) \tag{18}$$

where $\boldsymbol{N}(R, \bar{\mu}_{x_{max}}, \bar{\sigma}^2_{x_{max}})$ again denotes a univariate GP posterior (derived using Eq. 3 and 4), defined at a single prediction location, now equal to x_{max}.

D. An illustrative example

This section aims to demonstrate the active learning rules, and reconsiders the illustrations provided in **Fig. 1** in more detail. Data for this example were simulated, and are described in Section V. The initial stimulus level was set to $x_1 = 70$ dB HL, and the targets for the GP were set to $T_1 = 0.5$, $T_2 = 0.3$, $T_3 = 0.2$, $T_4 = 0.15$ and $T_5 = 0.1$ μV with corresponding δ_i thresholds of $\delta_1 = 0.3$, $\delta_2 = 0.2$, $\delta_3 = 0.1$, $\delta_4 = 0.075$, and $\delta_5 = 0.05$ μV, and $x_{max} = 90$ dB HL. The GP furthermore aimed to estimate $f(x)$ along the [-10, 90] dB HL interval with a 1 dB resolution, i.e., the $\boldsymbol{X}_P$ prediction locations comprised 101 integers ranging from -10 to 90. The GP was also allowed to test any level along this interval, but maintaining the 1 dB resolution. The mean and covariance functions were specified as described in Section III, and were used to construct the

GP prior, a simplified depiction of which is shown in **Fig. 1, panel (a)**. The simulated PTTa growth function (i.e., the $f(x)$ function) is also shown as a dotted line.

After specifying the GP prior, 500 epochs were collected (albeit simulated) at level $x_1 = 70$ dB HL. A coherent average was then computed along with a (biased) PTTa value of $a_1 = 0.92$ μV, which was replaced with an (unbiased) estimate of $o_1 = 0.62$ μV, as described in Section III.A. The standard deviation of o_1 was estimated to be $\sigma_1 = 0.26$ μV. The GP prior was thus updated using $\boldsymbol{O}_T = o_1 = 0.62$ and $\boldsymbol{\sigma}_T^2 = \sigma_1^2 = 0.26^2 = 0.0676$, giving the GP posterior shown in **Fig. 1, panel (b)**.

Next, the GP posterior was inspected to determine the next stimulus level. As explained previously, the GP's initial priority is to obtain a roughly monotonic estimate of $f(x)$, as this provides directional guidance when locating the T_i targets. To facilitate a monotonic estimate, the most likely PTTa at x_{max} was first estimated using Eq. (18), giving $R_{max} = 0.29$ μV. Since $R_{max} < T_1$, this suggests that the GP should collect data at x_{max} to facilitate a monotonic estimate. However, it is also important to consider the level of confidence associated with the R_{max} estimate. If the GP exhibits uncertainty regarding this estimate, then additional data at x_{max} is indeed deemed necessary, whereas if the GP is sufficiently confident regarding its estimate, then it is instead allowed to locate the next target. In this example, the standard deviation of the GP posterior at x_{max} was $\bar{\sigma}_{x_{max}} = 0.33$ μV, and thus $\bar{\sigma}_{x_{max}} > \delta_1 = 0.3$ μV, indicating too much uncertainty regarding the R_{max} estimate. The next stimulus level to test at was therefore set to $x_2 = x_{max} = 90$ dB HL.

An additional 500 epochs were thus simulated using a stimulus level of 90 dB HL, and an unbiased PTTa of $o_2 = 0.96$ μV was estimated along with a standard deviation of $\sigma_2 = 0.24$ μV. The MVN prior was again updated in accordance with prior assumptions and $\boldsymbol{O}_T = [o_1, o_2]$ and $\boldsymbol{\sigma}_T^2 = [\sigma_1^2, \sigma_2^2]$, giving the GP posterior shown in **Fig. 1, panel (c)**. The posterior was then inspected to determine the next stimulus level. The most likely PTTa at x_{max} was now $R_{max} = 0.73$ μV, thus exceeding the T_1 target of 0.5 μV, which suggests that a roughly monotonic estimate of $f(x)$ has now been obtained. The GP therefore proceeds to estimate the most likely location for T_1 using Eq. (17), giving $x_{T_1} = 67$ dB HL. However, the standard deviation of the GP posterior at x_{T_1} is $\bar{\sigma}_{x_{T_1}} = 0.21$ μV, i.e., the GP is already moderately confident regarding its $f(67)$ prediction, which suggests that the T_1 target might already be deemed located. Indeed, for this example, the $\bar{\sigma}_{x_{T_1}} < \delta_1 = 0.3$ μV condition was met. The GP therefore proceeds to estimate the $T_2 = 0.2$ μV target. The most likely x where $f(x) = T_2$ was $x_{T_2} = 54$ dB HL, and $\bar{\sigma}_{x_{T_2}} = 0.27$ μV. The $\bar{\sigma}_{x_{T_2}} < \delta_2 = 0.2$ μV condition was not met, which implies that the GP is not sufficiently confident regarding its $f(54)$ prediction, and hence that additional data collection at this level is necessary. The next stimulus level to test at was therefore $x_2 = x_{T_2} = 54$ dB HL.

Fast forwarding a bit, **Fig. 1, panel (d)** shows the GP posterior after 44 iterations, with each iteration incorporating an additional 500 epochs into the data set. It is worth noting that there are a total of 16 data points as some stimulus levels were revisited. The final x_{T_i} estimates were 52, 42, 38, 36, and 34 dB HL, for $i = 1, 2, 3, 4, 5$, respectively. The corresponding standard deviations of the GP posterior at these locations were 0.11, 0.071, 0.055, 0.049, and 0.047 μV, respectively. The stopping criterion was thus met for all i and data collection was stopped. Finally, hearing threshold was estimated using Eq. (17) using $T_i = 0$, giving $x_{HT} = 29$ dB HL.

IV. A SEQUENTIAL HOTELLING'S T^2 TEST TO COMPARE AGAINST

When evaluating the performance of the GP, it is important to establish a benchmark to compare against. The HT^2 test was therefore also included in the assessment, which previously outperformed various alternative methods for ABR detection [19]. However, as mentioned in the introduction, fully automated ABR hearing threshold estimation also requires a sequential test strategy for determining when to analyse data and for controlling the error rates, along with a stimulus selection protocol for choosing the next level and homing in on hearing threshold. These components are described in the sections below.

A. The test statistic

When used for ABR detection, the HT^2 test evaluates the hypothesis that the expected value for the coherently averaged epoch is zero. More specifically, the null hypothesis is defiend as $H_0 : \bar{v}_i = 0$ for $i = 1, 2, ..., Q$, where $\bar{v}_i$ denotes the i^{th} "mean voltage mean". As the name suggests, a voltage mean is defined as a mean voltage value, taken across a short interval of EEG data. In particular, each 0-15 ms epoch was compressed into $Q = 25$ voltage means by averaging across 0.6 ms intervals [42], giving a G_n x Q dimensional matrix of features, say $\boldsymbol{V}$, where G_n is the number of epochs recorded at stimulus level x_n. The $\bar{v}_i$ values are then computed by averaging down the Q columns of $\boldsymbol{V}$. The T^2 statistic itself is given by [50]:

$$T^2 = G_n[\bar{v}_1, \bar{v}_2, ..., \bar{v}_Q]\boldsymbol{S}^{-1}[\bar{v}_1, \bar{v}_2, ..., \bar{v}_Q]' \qquad (19)$$

where $\boldsymbol{S}^{-1}$ is the inverse of the covariance matrix of $\boldsymbol{V}$, and $'$ denotes transpose. The T^2 statistic can be transformed into an F-statistic using $\frac{T^2(G_n - Q)}{Q(G_n - 1)}$, which is F-distributed with Q and $G_n - Q$ degrees of freedom under H_0 [50].

B. The sequential test strategy

In practice, the accuring ABR data is usually analysed repeatedly, as this allows data collected to be stopped as soon as an ABR is deemed present or absent, thus keeping test time low. The caveat is that the probability of false positives (i.e. detecting an ABR when none is present) increases with the number of hypothesis tests carried out [46]. Various sequential test procedures have therefore been proposed, which aim to control the FPR by carefully constructing the critical thresholds for response detection [39-44]. In this study, the Convolutional Group Sequential Test (CGST) from [43] was adopted, which allows two-sided critical thresholds to be constructed, i.e., thresholds for inferring both ABR present

and ABR absent. This is important for hearing threshold estimation where inference is needed on both "ABR present" and "ABR absent". It is also worth mentioning that the CGST previously demonstrated reductions in test time of 40-45% relative to a conventional "single shot" test where data is analyzed just once [42]. However, extensive comparisons between sequential ABR detection methods are lacking in the literature, and the CGST may not be the optimal approach.

1) The Convolutional Group Sequential Test: When using the CGST, data is analyzed in disjoint blocks of observations. For example, when analysing 6000 epochs with a 3-staged sequential test, epochs 1-2000 might be analysed at stage one, epochs 2001-4000 at stage two, and epochs 4001-6000 at stage three. Data analyzed in previous stages thus cannot be re-analyzed in subsequent stages as the CGST assumes independence between stages.

At each stage, a p value is computed by the HT^2 test. These p values are potentially transformed, and then summed with p values from previous stages. In the current study, p values were log-transformed and combined using Fisher's method, which was chosen as it previously outperformed various alternative combination functions in terms of Bahadur efficiency [51]. The log-transformed sum of p values is henceforth referred to as the "summary statistic" and is defined as:

$$S_k = \sum_{k=1}^{K} -2\log(p_k) \tag{20}$$

where p_k is the p value generated at stage k. After each stage, a decision can be made regarding the presence or absence of an ABR: If $S_k > C_k$, then H_0 is rejected and "ABR present" is concluded, whereas if $S_k < B_k$, then H_0 is accepted and "ABR absent" is concluded. If neither condition is met, then the trial proceeds to the next stage, up to a maximum of K stages.

The main challenge is to find the C_k and B_k critical thresholds for controlling the stage-wise FPRs and true-negative rates (TNRs). The stage-wise FPRs are denoted by α_1, α_2, ... , α_K, and are chosen freely by the user at the outset, albeit under the condition that $\sum_{k=1}^{K} \alpha_k = \alpha$ where α is the desired FPR for the full sequential test. The stage-wise TNRs are denoted by β_1, β_2, ... , β_K, and are also chosen by the user at the outset, under the condition that $\sum_{k=1}^{K} \beta_k = \beta$ where β is the TNR for the full sequential test.

The CGST thus aims to find C_k and B_k (for $k = 1, 2, ..., K$), such that stage-wise FPRs equal α_1, α_2, ... , α_K, and the stage-wise TNRs equal β_1, β_2, ... , β_K. The approach builds on work from [46], and revolves around numerically convolving truncated probability density functions. A comprehensive description of the full procedure is outside the scope of the current work, but a detailed overview is given in [43] and [42]. What is important to note, however, is that several parameters need to be specified at the outset, including the total number of stages K, along with the α_k and β_k values for $k = 1, 2, ..., K$. The stage-wise ensemble sizes, denoted by N_k, are typically also pre-specified.

In the current study, the number of stages was set to $K = 5$, and the α_k and β_k parameters were set to $\alpha_k = \frac{0.01}{K}$ and $\beta_k = \frac{1-0.01}{K}$ for all k, chosen based on results from [42]. This led to the stage-wise critical thresholds presented in Table I. The N_k values were furthermore set to 2000, giving a maximum ensemble size of 10,000 epochs after $K = 5$ stages, which was determined based on pilot simulations (details not presented).

TABLE II
THE STAGE-WISE CRITICAL THRESHOLDS FOR THE SEQUENTIAL HOTELLING'S T^2 TEST.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
C_k	12.429	16.011	19.111	21.844	24.552
B_k	0.441	2.347	5.444	10.059	24.552

C. The stimulus selection protocol

In order to home in on hearing threshold with the sequential HT^2 test, the stimulus level was adjusted in ± 10 dB steps, i.e., a 10-down-10-up approach was adopted. Specifically, if H_0 was rejected and an ABR was deemed present, then the stimulus level was decreased by 10 dB, whereas if H_0 was accepted and an ABR was deemed absent, then the stimulus level was decreased by 10 dB. Data collection was stopped once H_0 was both accepted and rejected, after which hearing threshold was estimated by taking the average of the highest H_0 acceptance level and the lowest H_0 rejection level.

V. EVALUATING TEST PERFORMANCE IN SIMULATIONS

This section evaluates the performance of the GP in simulated data and draws comparisons with the sequential HT^2 test.

A. Data

The simulated data comprised (1) an ABR template for simulating a response, (2) coloured noise for emulating the EEG background activity, and (3) sinusoid ramps for simulating the PTTa growth functions.

1) ABR template for simulating a response: The template for simulating an ABR was created by averaging ABR measurements from 12 normal-hearing adults. The subject data comprised a 33.33 Hz click-evoked ABR threshold series, previously described in [18]. Prior to constructing the template, data were band-pass filtered from 30-1500 Hz using a 6^{th}-order Butterworth filter, and artefact rejection was applied using a ± 10 μV threshold. All coherent averages were inspected visually to confirm that a clear ABR was present. The resulting ABR template is shown in **Fig. 3, panel (a)**. It is worth noting that that averaging measurements from multiple test subjects may have attenuated the high-frequency content in the template waveform, effectively simulating a low pass filter.

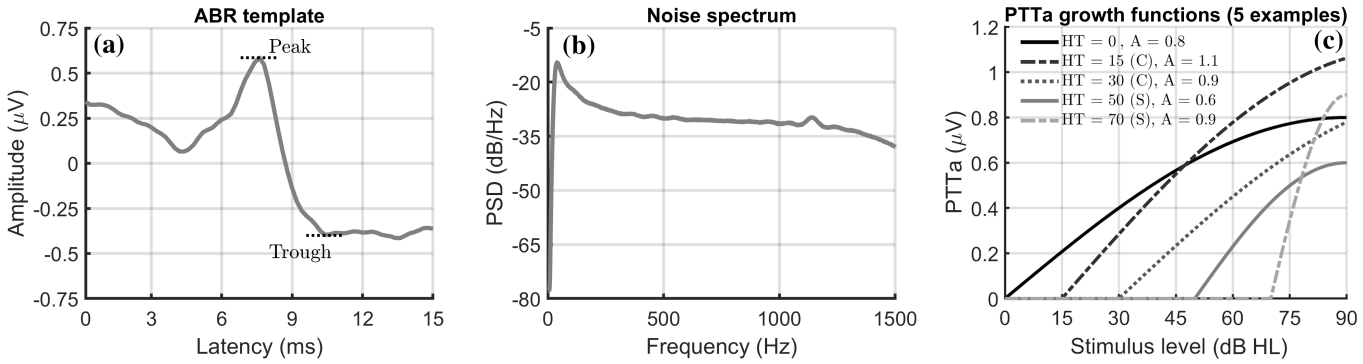

Fig. 3. The models for generating the simulated data. **Panel (a)** shows the ABR template for simulating a response, which was constructed by averaging ABR measurements from multiple subjects. **Panel (b)** shows an example of the power spectral density (PSD) function that underlies the colored noise, emulating the EEG background activity. **Panel (c)** shows 5 examples of simulated PTTa growth functions, generated using sinusoidal ramps (equations 21 and 22). For these particular PTTa growth functions, the hearing threshold (HT) was set to either 0, 15, 30, 50 or 70 dB HL, and the maximum amplitude at saturation (denoted by "A" in the legend) was set to 0.8, 1.1, 0.9, 0.6 or 0.9 μV. In two cases, a conductive hearing loss was simulated (denoted by "C") and in two cases a sensorineural hearing loss (denoted by "S").

2) Coloured noise for simulating the background activity: Coloured noise with spectral content similar to subject-recorded EEG was generated by filtering Gaussian white noise with all-pole filters, where the poles of the filters were given by the parameters of 60^{th}-order auto-regressive (AR) models. The AR models were estimated from recordings of EEG background activity (no stimulus used) from [13] with a new model being fit to each recording. Prior to fitting the AR models, data were again band-pass filtered from 30-1500 Hz using a 6^{th}-order Butterworth filter, and artefact rejection was applied using a ± 10 μV threshold. The resulting colored noise was also band-pass filtered from 30-1500 Hz. This resulted in SNRs ranging from approximately -36 dB to -20 dB, depending on the amplitude of the ABR template as well as variations in the simulated noise. An example of the spectrum underlying randomly generated noise is shown in **Fig. 3, panel (b)**.

3) Sinusoidal ramps for simulating PTTa growth functions: PTTa growth functions were simulated using sinusoidal ramps. The aim was to generate a wide range of growth functions that represent various degrees of sensorineural and conductive hearing loss. Each ramp was defined along the [0, 0.5π] interval, which was assumed to correspond to stimulus levels along the [-10, 90] dB HL interval. In particular, PTTa growth functions were simulated using:

$$f(x) = A \cdot sin(\acute{x}) \tag{21}$$

where A is the PTTa in μV at saturation, representing the largest PTTa value along the growth curve, and $\acute{x}$ represents the stimulus level, but scaled to the [0, 0.5π] interval. The latter was computed using:

$$\acute{x} = r \cdot max\{0, x - x_{HT}\} \tag{22}$$

where x_{HT} represents the hearing threshold and r is the stepwise change along the [0, 0.5π] interval per unit increase in level x. When emulating a conductive hearing loss, $r = \frac{0.5\pi}{90}$, whereas for a sensorineural hearing loss, $r = \frac{0.5\pi}{90 - x_{HT}}$. Note that the sensorineural hearing loss compresses the ramp to the [x_{HT}, 90] dB HL interval, whereas the conductive hearing loss shifts the ramp back and forth along x based on x_{HT}. These choices were motivated by findings from the literature, which show similar trends in PTTa growth functions for these types of hearing loss [12]. Some examples of simulated PTTa growth functions are shown in **Fig. 3, panel (c)**.

B. Test conditions

Sequential HT^2 test: The performance of the sequential HT^2 test was evaluated for a range of stimulus starting levels. As the name suggests, the starting level refers to the initial stimulus level at which testing begins, which was varied from 0 to 90 dB HL, in steps of 10 dB HL. The aim was to investigate the impact of the starting level within the "10-down-10-up" protocol, but also to ensure that the GP was not compared to a sub-optimal sequential HT^2 test.

For each starting level, a total of 10,000 hearing threshold estimation trials were simulated. In each simulated trial, a conductive or a sensorineural hearing loss was randomly chosen, each with a 50% chance, and both x_{HT} and A were randomly resampled from uniform distributions, defined along the [0, 70] dB HL and [0.75, 1.25] μV intervals, respectively.

GP approach: For the GP approach, the T_i target values were specified as $T_1 = 0.5$, $T_2 = 0.3$, $T_3 = 0.2$, $T_4 = 0.15$, and $T_5 = 0.1$ μV, and the corresponding δ_i threshold values were set to $\delta_1 = 0.2$ and $\delta_2 = \delta_3 = \delta_4 = 0.1$. Test performance was then evaluated for different δ_5 choices, which was varied from 0.03 to 0.1 μV in increments of 0.005. The δ_5 threshold was varied as it was assumed to be the most impactful on test performance. In particular, if the predictions for $f(x) = 0.1$ μV are accurate, then extrapolating down to $f(x) = 0$ μV will presumably also yield an accurate hearing threshold estimate. For each δ_5 value, a total of 10,000 hearing threshold estimation trials were simulated following the same procedure as described previously for the Sequential HT^2 test.

The starting level was always set to 70 dB HL and data were analysed every 500 epochs.

C. Results

This section presents the simulation results for the Sequential HT^2 test and the GP approach. Test performance was evaluated in terms of test time, as well as the "hearing threshold estimation error", defined as the estimated hearing threshold minus the true simulated hearing threshold.

1) Sequential HT^2 test: Results are presented in **Fig. 4, panels (a)** and **(b)**. Panel (a) shows the median test time, along with the 0.68, 0.95, and 0.99 quantiles of the test times, plotted as a function of the starting level. Panel (b) shows the hearing threshold estimation errors, presented in the same format as panel (a).

For the sequential HT^2 test, the best test performances were observed when initiating the 10-down-10-up approach at a relatively high stimulus level of approximately 60 dB HL. In particular, test times were shortest when starting the test at around 50 or 60 dB HL, and test accuracies were highest when starting the test at 50 dB HL and above. The longer test times for the lower starting levels indicate that more data was generally needed to determine the absence of an ABR compared to determining the presence of an ABR. With respect to the reduced test accuracies for lower starting levels, this is due to false-positives. In approximately 1% of the cases, false-positives led to estimation errors of over 45 dB.

2) The GP approach: Results are presented in **Fig. 4**, **panels (c)** and **(d)**. Panel (c) shows the median test time, along with the 0.68, 0.95, and 0.99 quantiles of the test times, now plotted as a function of the δ_5 threshold. Panel (d) shows the hearing threshold estimation errors, presented in the same format as panel (c).

As expected, both test time and test accuracy increased as more stringent stopping criteria (smaller δ_5) were used. Contrary to the sequential HT^2 test, which showed median estimation errors of 17 dB, the GP was more or less unbiased, i.e. the median estimation error was $\sim$0 for all δ_5 values. The spread of the errors, however, increased with the δ_5 stopping criterion. Assuming a hearing threshold estimation error of approximately $\pm$10 dB is acceptable, then a suitable choice for δ_5 would be $\sim$0.035 μV, which led to 99% error quantiles of [-12, 7] dB and a median test time of $\sim$7 minutes.

3) Comparing methods: To obtain a fair comparison between methods, it is helpful to equate their test times and compare their errors, or vice versa. The sequential HT^2 with a starting level of 60 dB HL had a median test time of $\sim$7 minutes along with 99% error quantiles of [4, 32] dB, or a 28 dB error range. For the GP, a similar $\sim$7 minute median test time was obtained when using $\delta_5 = 0.035$ μV, which led to 99% error quantiles of [-12, 7] dB, corresponding to a 19 dB error range. The GP thus demonstrated a reduction of 9 dB in the error range compared to the sequential HT^2 test while maintaining a similar test time. The GP furthermore obtained a $\sim$28 dB error range when using $\delta_5 = 0.055$ μV, which led to a median test time of approximately 3.5 minutes. This represents a roughly 50% reduction in median test time compared to the sequential HT^2 test while maintaining a similar error range.

VI. Case Study

This section aims to establish a proof of concept for the GP approach in subject data, and evaluates test performance in ABR recordings from an adult volunteer with normal hearing.

1) Data: ABRs were evoked using a 4 kHz chirp as stimulus. The stimulus was presented through ER-2 insert phones, and was calibrated using a Brüel and Kjaer type 2112 sound level meter with the output of the sound level meter routed to an oscilloscope to allow measurement of peak-to-peak amplitude. The peak-to-peak amplitude for a 94 dB SPL calibration piston was first measured as a reference point. Chirp calibration in dB HL was then carried out with reference to the 0 dB HL peak-to-peak amplitude values given in the International Organization for Standardization (ISO) 389-6: 2007 along with the UK National Hearing Screening Protocol (NHSP) recommended stimulus reference levels for ABRs (correction factors were applied to the stimuli output levels to obtain dB HL).

During the test, the subject reclined in a comfortable chair and was asked to relax with eyes closed. Chirp stimuli were generated in Matlab and routed via an RME Fireface Soundcard to ER2 insert phones placed in the subject's ears. The chirp stimulus was presented at a rate of 47.17 Hz at a range of dB sensation levels (SLs), i.e., relative to subjects' behavioural hearing threshold. The behavioural hearing threshold was estimated to be 5 dB HL, and was found using a standard 10-down-5-up test procedure, i.e., the chirp intensity was decreased in steps of 10 dB for every correct response, and increased by 5 dB for every missed response. The chirp was then presented from -20 to 50 dB SL, in steps of 10 dB SL, corresponding to dB HLs ranging from -15 to 55 dB HL. ABR measurements were recorded at 48 kHz using an Interacoustics Eclipse system with electrodes placed at the vertex (active electrode), the nape of the neck (reference) and mid-forehead (ground). Electrode impedances remained below 5 kΩ throughout the recording. The Eclipse data was routed back to Matlab via the RME Fireface where it was band-pass filtered from 30-1500 Hz using a 6^{th}-order Butterworth filter. Artefact rejection was also applied using a $\pm$20 μV rejection level. Approximately 10,000 artefact-free epochs were recorded at each stimulus level. This study was approved by the Faculty Ethics Committee at the University of Southampton (ERGO II 56025.A3).

2) Analysis: The GP was applied offline to the recorded ABR data, but emulating the online data collection procedure. The GP's starting level was set to 55 dB HL, and data were analysed every 500 epochs. Note that the GP's test locations were limited to stimulus levels that were available in data, i.e., -15 to 55 dB HLs in 10 dB HL increments. The GP's prior mean and covariance functions were specified as described in

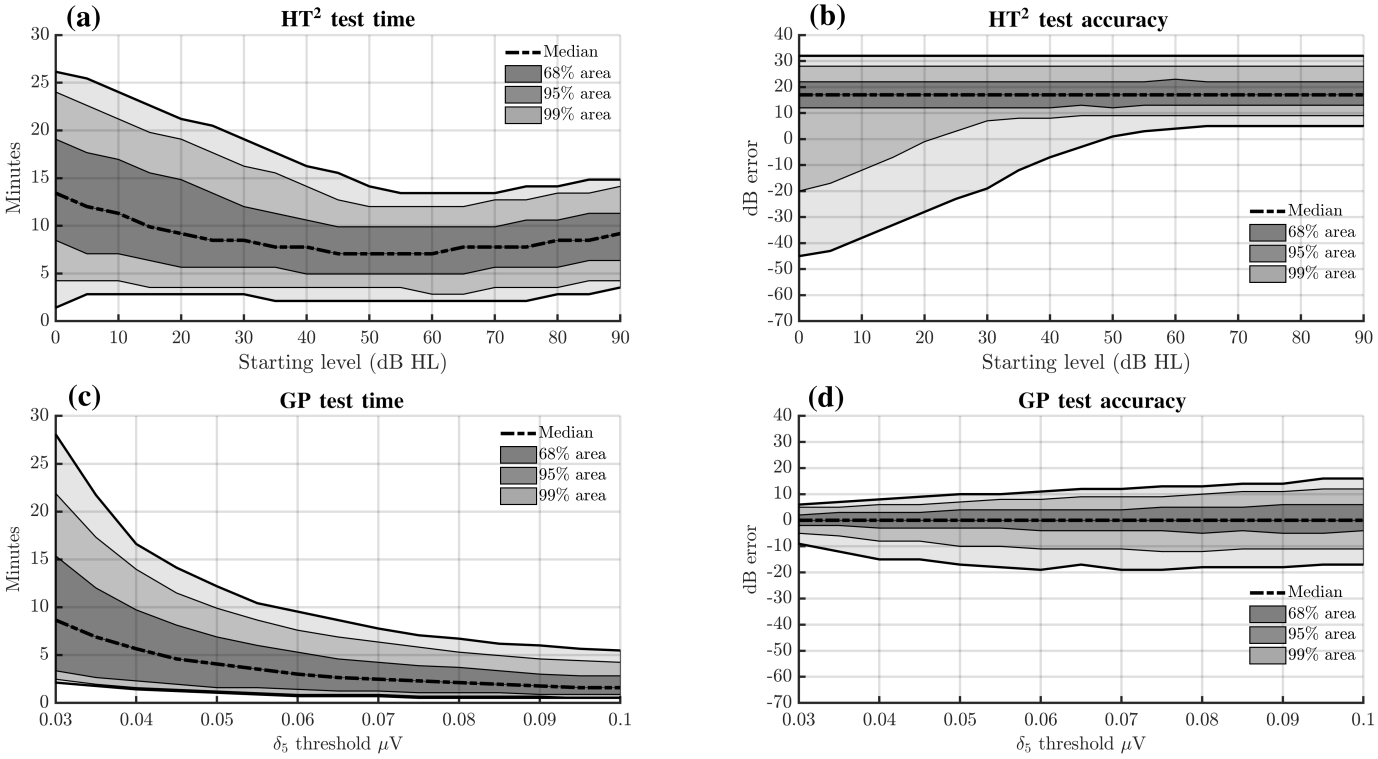

Fig. 4. Results from the simulations. **Panel (a)** shows the test times for the sequential Hotelling's T^2 test, plotted as a function of the initial stimulus level. The thick dashed lines show the median values, and the shaded areas indicate the regions where 68%, 95%, and 99% of the observed values fell. **Panel (b)** shows the test accuracies for the sequential Hotelling's T^2 test. Test accuracy is given by the "dB error," defined as the estimated hearing threshold minus the true hearing threshold. The thick dashed line represents the median dB error, and the shaded areas again represent the regions where 68%, 95%, and 99% of the observed values fell. **Panels (c)** and **(d)** show the test times and dB errors for the GP, shown in the same format as panels (a) and (b), presented as a function of the $\boldsymbol{\delta_5}$ stopping criterion. The $\boldsymbol{\delta_5}$ threshold represents the required level of certainty in the GP's predicted values before data collection can be stopped, with smaller $\boldsymbol{\delta_5}$ values indicating more stringent requirements for the GP's predictions.

Section III, and the targets for the GP were set to $T_1 = 0.5$, $T_2 = 0.3$, $T_3 = 0.2$, $T_4 = 0.15$, and $T_5 = 0.1$ μV, with corresponding δ_i thresholds of $\delta_1 = 0.2$, $\delta_2 = 0.1$, $\delta_3 = 0.1$, $\delta_4 = 0.1$, and $\delta_5 = 0.05$ μV.

In order to establish a rough benchmark to compare against, data were also inspected visually by an examiner who followed guidelines provided by the British Society of Audiology [5]. Although developed for infant testing, rather than adults, these guidelines provide a rigorous set of rules for response detection, and thus help to mitigate examiner bias and ensure accurate test outcomes. The examiner initiated the test at 55 dB HL, and inspected the accruing data in increments of 500 epochs until an ABR was deemed present or absent. If an ABR was deemed present, the stimulus level was reduced by 10 dB HL. This repeated until an ABR was deemed absent, after which hearing threshold was assumed to be located at the lowest stimulus level where an ABR was detected.

3) Results: The GP's predictions for the PTTa growth function are presented in **Fig. 5**: **Panel (a)** shows a simplified depiction of the GP prior, representing the GP's predictions before observing data. **Panel (b)** then shows the GP posterior after having located $T_1 = 0.5$ μV, **Panel (c)** after having located $T_2 = 0.4$, $T_3 = 0.2$, and $T_4 = 0.15$ μV, and **panel (d)** shows the GP posterior after all targets were located. The final GP-estimated hearing threshold was 5 dB HL, and was equal to the behavioural hearing threshold, indicating an estimation error of 0 dB. The total test time was furthermore 360.4 seconds, or ∼6 minutes.

With respect to the visual inspection results, the examiner concluded "ABR present" at stimulus levels 55, 45, 35, 25, and 15 dB HL, and concluded "ABR absent" at 5 dB HL. The ABR waveforms inspected by the examiner are shown in **Fig. 6**. Assuming hearing threshold is located between the lowest level that an ABR was deemed present and the highest level that an ABR was deemed absent, then the estimated hearing threshold was 10 dB HL, i.e., 5 dB HL above the behavioural threshold. However, test time for the audiologist was 741.7 seconds, or ∼12.6 minutes. In this test subject, the GP thus demonstrated a slightly higher test accuracy whilst also reducing test time by approximately 50%. In future work, a more extensive assessment will be carried out to more thoroughly evaluate the GP's performance.

VII. Discussion

The GP is a flexible and potentially powerful nonparametric approach for smoothing, interpolation, and pattern discovery [49, 50, 52]. It was originally established in the field of geostatistics [53], and has since been applied in many areas,

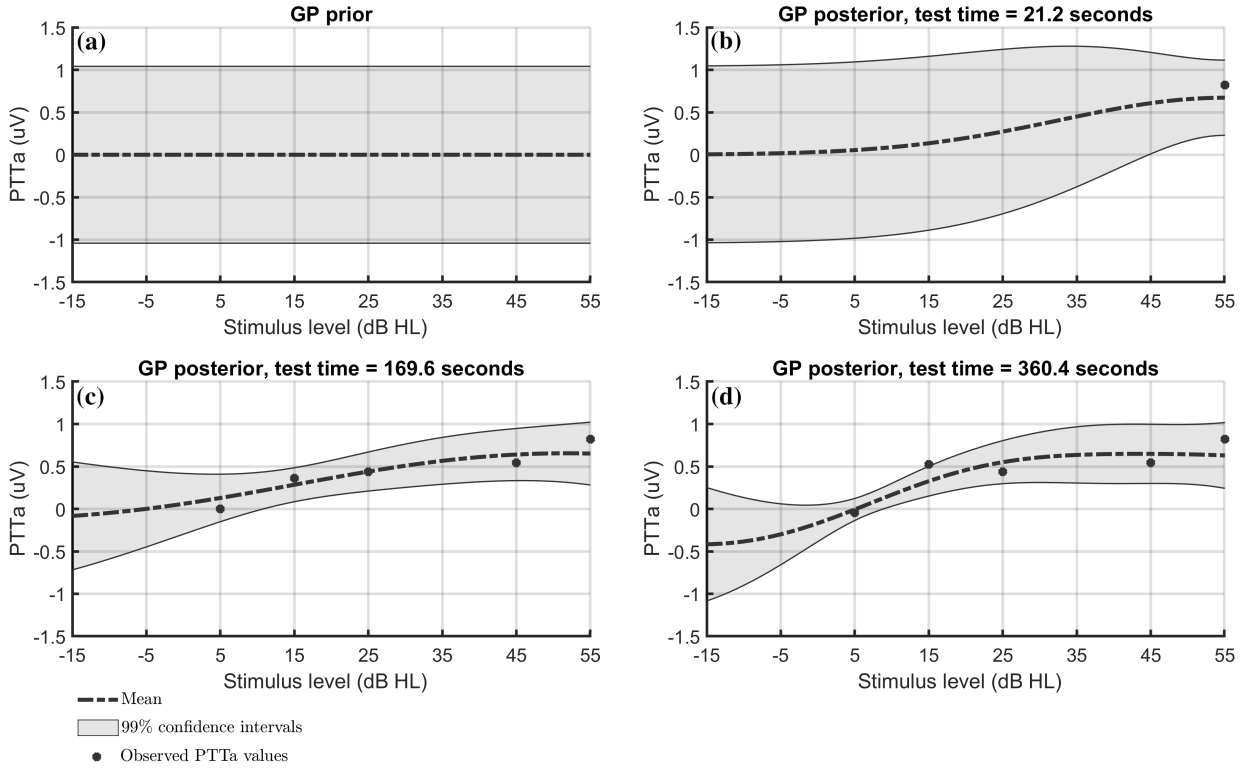

Fig. 5. The GP's predictions for the PTTa growth function for an adult volunteer with normal hearing. **Panel (a)** shows the GP prior, representing the GP's predictions before having observed data. **Panel (b)** shows the GP posterior after having observed data at 55 dB HL. At this point in the test, the GP has already located its first target, equal to $\boldsymbol{T_1 = 0.5}$ μV. **Panel (c)** shows the GP posterior after having observed additional data at 45, 25, 15 and 5 dB HL. The GP has now also located targets $\boldsymbol{T_2 = 0.4}$, $\boldsymbol{T_3 = 0.2}$, and $\boldsymbol{T_4 = 0.15}$ μV. Finally, **panel (d)** shows the GP posterior after all targets were located. The estimated hearing threshold for this subject was 5 dB HL, and was equal to the behavioural hearing threshold. The total test time was ∼6 minutes, which was relatively long as data were noisy. Results are further considered in the main text.

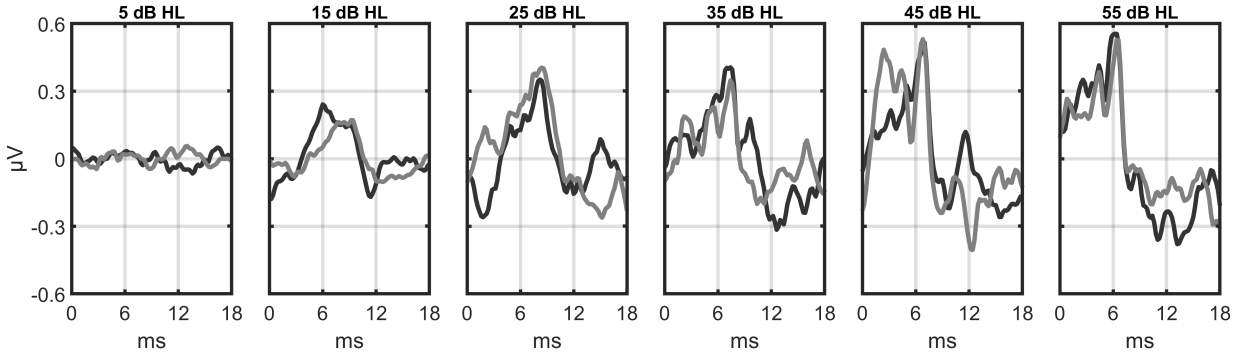

Fig. 6. Auditory Brainstem Response (ABR) waveforms that were visually inspected by an examiner when estimating hearing threshold in an adult volunteer with normal hearing. An ABR was deemed present at 55, 45, 35, 25, and 15 dB HL, and absent at 5 dB HL. Hearing threshold was assumed to be located between 5 and 15 dB HL, giving an estimated threshold of 10 dB HL, which closely aligned with the behavioural hearing threshold of 5 dB HL.

including finance [54], astronomy [55], genomics [56], imaging [57], epidemiology [58] and general optimisation [59], just to name a few areas. The GP has recently also shown potential in the related field of behavioural hearing testing where it was used to estimate subjects' audiogram using behavioural Pure Tone Audiometry [60]. Results from [60] show rapid approximations of the audiogram, in 4 minutes or less, which contrasts with the 15-20 minutes for the standard clinical protocol, i.e. a 70 to 80% reduction in test time was obtained. A similar test time reduction would be transformative for ABR measurements, and was the primary incentive for exploring the GP approach in the current work.

Results from the current study show reductions in test time of ∼50% for the GP relative to the sequential HT^2 test. The reduction in test time is substantial, but still smaller than that observed for behavioural audiogram estimation in [60], which begs the question as to whether further test time reductions can be obtained. An important distinction between the current study and [60] is that the GP in [60] considered correlations across both stimulus levels and frequencies, whereas the GP in the current work considered correlations across stimulus levels only. Although both studies differ substantially in data

and methodology, and thus cannot be compared directly, it is envisaged that further reductions in test time might be obtained by also incorporating correlations across stimulus frequencies.

Additional reductions in test time might also be obtained by considering all peaks and troughs within the ABR, as opposed to just the wave V PTTa. The correlation structure underlying these peaks and troughs is, however, intricate, with some peak-trough complexes shifting in latency and/or appearing/disappearing as a function of the stimulus level [1]. In the current work, it was therefore opted to simplify the approach, and to compress the ABR into a single PTTa value, effectively discarding the peak-trough correlation structure. This may have contributed towards a sub-optimal test performance.

On a related note, the PTTa statistic is known to be less sensitive than some other ABR test statistics (e.g. [18]), which suggests that it may be preferable to apply the GP to these alternative, more powerful statistics. Unfortunately, the assumptions underlying the GP are grossly violated for most ABR test statistics, the T^2 statistic included. These violations are essentially the same as (but more severe than) those described in Section III for the PTTa. The T^2 statistic, for example, increases indefinitely with the SNR. This is of course not to say that the GP cannot be applied to other ABR test statistics, but that work is needed to find suitable data transformations, such that the data conforms to the GP's assumptions.

With respect to the active learning rules, these were designed to locate various non-zero PTTa values along $f(x)$, after which the GP posterior was used to extrapolate down towards $f(x) = 0$. As discussed previously, locating the larger PTTa targets helps to facilitate a monotonic estimate of $f(x)$, which then provides directional guidance on where the remaining targets, as well as hearing threshold, are located. This, however, raises the question as to how the PTTa targets should be chosen. One factor to consider is the distance between hearing threshold and the smallest T_i target, which determines the distance across which the GP has to extrapolate across when estimating $f(x) = 0$. When this distance is large, relatively small random errors could be inflated, leading to relatively large hearing threshold estimation errors. In future work, the choice for the PTTa targets may need to be tweaked and/or optimised for a cohort of test subjects, and ideally for the main target population, i.e. infants with suspected hearing loss.

An alternative strategy is to not extrapolate towards $f(x) = 0$, but to simply assume that hearing threshold is located at a fixed distance below some minimum PTTa target, say T_{min}. For example, it might be assumed that hearing threshold is 10 dB lower than the x value where $f(x) = T_{min} = 0.1$ μV. The accuracy of this approach depends on inter-subject variations in the dB difference between hearing threshold and the x value where $f(x) = T_{min}$. A rough indication of this variance might be found in the literature: In [11], PTTa values were $\sim$0.16 μV for 10 dB nHL clicks, and $\sim$0.05 μV for 10 dB nHL tone pips, whereas in [12], PTTa values were $\sim$0.25 μV for 10 dB nHL 1 kHz tones. This suggests a moderate amount of variation in PTTa values evoked by 10 dB nHL stimuli, which might be due to variations in individual PTTa growth functions, or differences in stimulus parameters, recording conditions, and/or data analysis techniques, which impedes direct comparisons across studies. It is also worth noting that the reported PTTa values might be over-estimated due to the PTTa bias, but potentially also under-estimated as they represent group averages, i.e. it is feasible that ABRs were not present in all test subjects, especially at low levels, which may have resulted in under-estimated mean PTTa values close to hearing threshold.

With respect to the choice for the prior, this can have a large impact on the performance of the GP, and should be chosen carefully, especially the covariance function [49]. Starting with the prior mean, this was defined under the assumption of "ABR absent", and was set to zero for all stimulus levels. Doing so facilitates monotonic estimates of $f(x)$, and ultimately helps to provide directional guidance on hearing threshold location. It is worth noting that additional simulations were carried out using non-zero prior means (details not presented), which often led to non-monotonic estimates of $f(x)$, and in less efficient decision making when adjusting the stimulus levels.

As for the covariance function, this was specified using the exponential kernel: the scale parameter s was chosen based on prior knowledge from the literature [11,12], whereas the θ parameter was treated as a random variable and was estimated (per recording) using MLE. The θ values were also confined to the [1000, 2000] interval, chosen based on pilot simulations. As noted in [29] and [45], θ relates to the number of expected zero-crossings (or changes in direction, i.e., from increasing to decreasing, or vice versa) through $\frac{1}{\theta^{-0.5}}$. The 1000 and 2000 boundaries thus correspond to a change in direction along $f(x)$ every $\frac{1}{1000^{-0.5}}$ = 31.6 dB HL and $\frac{1}{5000^{-0.5}}$ = 70.7 dB HL, respectively. These values were motivated by pilot simulations, and visual inspection suggests that the resulting function smoothness is roughly in line with PTTa growth function smoothness observed in the literature.

A. Comparisons with existing methods

The GP is an approach for conducting regression, and is therefore related to various "curve-fitting" methods, which have previously also been evaluated for ABR hearing threshold estimation [26-28]. Similar to the GP in the current work, these methods aim to estimate some form of the ABR's amplitude-intensity growth function, from which hearing threshold is ultimately inferred. In [27], the ABR's amplitude-intensity growth function was represented by a "root-mean-square growth function", and was estimated using sigmoid functions, whereas in [28], the amplitude-intensity growth function was represented by a "cross-correlation coefficient growth curve", and was estimated using power functions and sigmoid functions. Finally, in [26], a self-supervised random forest regression model was used to predict sound intensity levels, followed by the fitting of a piece-wise function consisting of a constant element and a 4th order polynomial.

One difference between the curve-fitting methods from [26-28] and the GP, is that methods from [26-28] assume a specific functional form for $f(x)$. If this assumption does not hold, then the accuracy of the estimated hearing thresholds may suffer.

This contrasts with the GP, which does not confine $f(x)$ to a specific functional form, but instead imposes smoothness constraints on the $f(x)$ function values. This offers greater flexibility in estimating $f(x)$, and allows the GP to converge on a wide range PTTa growth functions. This is important, as PTTa growth functions can vary considerably across individuals due to differences in hearing loss characteristics [1], e.g., depending on whether subject has a conductive or a sensorineural hearing loss. However, this flexibility comes at a cost, as the GP now needs to consider a wide range of potential $f(x)$ function values, which introduces additional uncertainty in the GP's estimates. In short, it is desirable to maximally constrain the estimates for the $f(x)$ function values, while still maintaining sufficient flexibility to converge on all possible PTTa growth functions. Whether the GP is the optimal approach to achieve this remains to be seen in future work.

A second difference between the curve-fitting methods from [26-28] and the GP, is that the GP was applied sequentially to the accruing data, whereas methods from [26-28] were applied in a post-hoc analysis. Sequential data analysis is an important component for ABR-related applications in the clinic, as it provides prompt feedback to clinicians, and generally helps to reduce test time. However, as mentioned previously, the risk is that sequential testing inflates the FPR, and adjusted critical decision boundaries are required to control the signficance level of the test.

B. False-positives

In the current work, FPRs for the sequential HT^2 test were controlled using the CGST approach [43]. It should be stressed, however, that while the CGST controls the FPR at each stimulus level, it does not prevent false-positives from occurring entirely. Results from **Fig. 4, panel (b)** indicate that the impact of these false-positives can be severe, in some cases leading to hearing threshold estimation errors of over 45 dB. This is concerning, especially when diagnosing and treating hearing loss in new borns where errors can lead to undiagnosed hearing loss, or worse, hearing damage due to over-amplified hearing aid settings. Although these large errors are rare, the severity of the repurcussions may be why fully automated ABR hearing threshold estimation are not yet trusted for fully autonomous use in the clinic, and supervision by an experienced clinician is still necessary to ensure accurate diagnoses and effective treatment of hearing loss.

To advance automated ABR hearing threshold estimation in the clinic, it is thus important to not only reduce test time, but to also ensure that large estimation errors do not occur. One way to mitigate these errors is to consider data from multiple stimulus levels simultaneously, and to exploit the correlations across ABR waveforms. By doing so, the impact of any single false-positive on the final estimated hearing threshold is increasingly "diluted" as data accrues across levels. This is likely also why large estimation errors were less common for the GP approach than for the sequential HT^2 test.

C. Limitations and future work

Perhaps the main limitation of the current study is that the GP approach was evaluated using simulations, with only an illustrative example in ABR data from an adult volunteer. As mentioned previously, simulated data is attractive as it allows large data sets to be constructed under controlled test conditions, i.e., data sets where both hearing thresholds and PTTa growth functions are known. However, there may be substantial differences between the simulated data and data encountered in the clinic. The extent to which results in this study generalize to clinical settings thus remains to be determined. Work is currently underway to further evaluate the GP approach in a cohort of normal-hearing and hearing-impaired subjects.

Another limitation of the current study is related to the active learning rules, which were not thoroughly optimized. This includes the choice of the T_i targets and δ_i thresholds, but also the concept of first estimating various non-zero regions along the PTTa growth function, which might not be the most efficient approach. Numerous active learning rule sets can be envisaged, and it is highly probable that the rules employed in this study are sub-optimal. Optimized active learning rules may also depend on the population being tested, and should be an important topic for future studies.

Further limitations are related to the GP itself, which does not consider the monotonicty property of the PTTa growth function. This property holds valuable information, and integrating it into the estimation procedure may help to further reduce uncertainty regarding the $f(x)$ estimates. Leveraging the monotonicity property may also help to mitigate the impact of false-positives and/or data outliers, which were observed to occasionally degrade hearing threshold estimation accuracy.

An additional limitation for the GP is that it assumes a single covariance structure for all PTTa values, i.e., it assumes that the rate at which PTTa values change is constant across all stimulus levels, which is not the case. For example, PTTa values are zero for all inaudible stimulus levels, and are thus fully correlated. Similarly, for very loud sounds, the PTTa value may saturate, also leading to highly correlated values. Sounds that transition from inaudible to moderately loud, on the other hand, may lead to relatively rapidly changes in PTTa values, and hence relatively weak correlation. In short, the rate at which PTTa values change is level-dependent, but this is not taken into account by the GP. This has implications when estimating the length scale parameter and may ultimately lead to over- or under-smoothing in the GP-estimated growth functions, and potentially reduced test accuracies and/or increased test times.

VIII. Conclusion

This work presented a GP approach with active learning rules for automatic ABR hearing threshold estimation. Simulation results show a $\sim$50% reduction in median test time for the GP relative to a sequentially applied Hotelling's T^2 test while maintaining similar test accuracy. In general, the GP is a flexible and potentially powerful approach for non-parameteric regression, but requires data to conform to its underlying assumptions. When applied to ABR PTTa data, computationally intensive data transformations using a bootstrap approach were needed to ensure that the GP's

assumptions were met. The GP approach in the current work also requires a suitable choice for test parameters, including the T_i targets, the δ_i thresholds, and the GP priors. In future work, these parameters should be evaluated and optimised in a large cohort of test subjects, and ideally for the main target population, i.e., infants with hearing loss. This work nevertheless suggest that the GP has much potential for improving objective ABR hearing threshold estimation, and results warrant further investigation of GPs in future work.

Repository

Matlab code for the Gaussian Processes is available at: https://github.com/mchesnaye/IEEE-Gaussian-processes-for-ABR-hearing-threshold-estimation-

Acknowledgements

This work was funded by the William Demant Foundation. The authors would also like to acknowledge the use of the IRIDIS High Performance Computing Facility and associated support services at the University of Southampton.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] T.W. Picton, *Human auditory evoked potentials*, San Diego, Plural Publishing Inc., 2011.

[2] D.L. Jewett and J.S. Williston, "Auditory-evoked far fields averaged from the scalp of humans," *Brain*, vol. 94, issue 4, pp. 681-96, Oct. 1971. doi: 10.1093/brain/94.4.681

[3] K. Hecox and R. Galambos, "Brain Stem Auditory Evoked Responses in Human Infants and Adults," *Arch Otolaryngol*, vol. 99, issue 1, pp. 30–33, Jan. 1974. doi: 10.1001/archotol.1974.00780030034006

[4] W.A. Selters and D.E. Brackmann, "Acoustic tumor detection with brain stem electric response audiometry," *Arch Otolaryngol*, vol. 103, issue 4, pp. 181-7, Apr. 1977. doi: 10.1001/archotol.1977.00780210037001

[5] G. Lightfoot, S. Brennan, J. FitzGerald, and I. Ferm, "Recommended Procedure, Auditory Brainstem Response (ABR) Testing in Babies," *The British Society of Audiology*. [Online]. Available: https://www.thebsa.org.uk [Accessed: Nov. 30, 2022]

[6] J.J. Eggermont, "Auditory brainstem response," in *Clinical Neurophysiology: Basis and Technical Aspects. Handbook of Clinical Neurology Series*, K. Levin and P. Chauvel, Eds. Elsevier, 2019, pp. 451-464. ISBN: 9780444640321.

[7] S.A. Wood, G.J. Sutton, and A.C. Davis, "Performance and characteristics of the Newborn Hearing Screening Programme in England: The first seven years," *Int J Audiol*, vol. 54, issue 6, pp. 353-358, 2015. doi: 10.3109/14992027.2014.989548

[8] D.R. Stapells and P. Oates, "Estimation of the pure-tone audiogram by the auditory brainstem response: a review," *Audiol Neurootol*, vol. 2, issue 5, pp. 257-80, Sep./Oct. 1997. doi: 10.1159/000259252 PMID: 9390836.

[9] Y.S. Sininger, L.L. Hunter, D. Hayes, P.A. Roush, and K.M. Uhler, "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 Hear*, vol. 39, issue 6, pp. 1207-1223, Nov./Dec. 2018. doi: 10.1097/AUD.0000000000000580

[10] R.M. Janssen, L. Usher, and D.R. Stapells, "The British Columbia's Children's Hospital tone-evoked auditory brainstem response protocol: how long do infants sleep and how much information can be obtained in one appointment?" *Ear Hear*, vol. 31, issue 5, pp. 722-4, Oct. 2010. doi: 10.1097/AUD.0b013e3181ddf5c0

[11] T.W. Picton, D.R. Stapells, and K.B. Campbell, "Auditory evoked potentials from the human cochlea and brainstem," *J Otolaryngol Suppl*, vol. 9, pp. 1-41, Aug. 1981. PMID: 7026799.

[12] J. Nousak and D. Stapells, "Auditory brainstem and middle latency responses to 1 kHz tones in noise-masked normally-hearing and sensorineurally hearing-impaired adults," *Int J Audiol*, vol. 44, issue 6, pp. 331-344, Jun. 2005. doi: 10.1080/14992020500060891

[13] S.M.K. Madsen, J.M. Harte, C. Elberling, and T. Dau, "Accuracy of averaged auditory brainstem response amplitude and latency estimates," *Int J Audiol*, vol. 57, issue 5, pp. 345-353, 2018. doi: 10.1080/14992027.2017.1381770

[14] M. Zaitoun, S. Cumming, A. Purcell, and K. O'Brien, "Inter and intra-reader variability in the threshold estimation of auditory brainstem response (ABR) results," *Hearing, Balance and Communication*, vol. 14, issue 1, 2016. doi: 10.3109/21695717.2016.1110957

[15] M. Vidler and D. Parkert, "Auditory brainstem response threshold estimation: Subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test," *Int J Audiol*, vol. 43, pp. 417-429, 2004. doi: 10.1080/14992020400050053

[16] C. Elberling and M. Don, "Quality estimation of averaged auditory brainstem responses," *Scandinavian audiology*, vol. 13, no. 3, pp. 187-197, 1984. doi: 10.3109/01050398409043059

[17] W. H. Martin, J. W. Schwegler, A. L. Gleeson, and Y-B. Shi, "New techniques of hearing assessment," *Otolaryngol. Clin. North Am.*, vol. 27, no. 3, pp. 487-510, 1994. doi: 10.1016/S0030-6665(20)30666-6

[18] J. Lv, D. M. Simpson, and S. L. Bell, "Objective detection of evoked potentials using a bootstrap technique," *Med Eng Phys*, vol. 29, no. 2, pp. 191-198, 2007. doi: 10.1016/j.medengphy.2006.03.001

[19] M. A. Chesnaye, S. L. Bell, J. M. Harte, and D. M. Simpson, "Objective measures for detecting the auditory brainstem response: comparisons of specificity, sensitivity and detection time," *Int J Audiol*, vol. 57, no. 6, pp. 468-478, 2018. doi: 10.1080/14992027.2018.1447697

[20] M. Romao and C. J. Tierra-Criollo, "A Bayesian approach to the spectral F-Test: Application to auditory steady-state responses," *Comput Methods Programs Biomed*, vol. 183, p. 105100, 2020. doi: 10.1016/j.cmpb.2019.105100

[21] R. M. McKearney, S. L. Bell, M. A. Chesnaye, and D. M. Simpson, "Auditory brainstem response detection using machine learning: a comparison with statistical detection methods," *Ear and Hearing*, vol. 43, no. 3, pp. 949-960, 2021. doi: 10.1097/AUD.0000000000001151

[22] M. A. Chesnaye, S. L. Bell, J. M. Harte, and D. M. Simpson, "Controlling test specificity for auditory evoked response detection using a frequency domain bootstrap," *J Neurosci Methods*, vol. 363, no. 1, p. 109352, 2021. doi: 10.1016/j.jneumeth.2021.109352

[23] E. Maris and R. Oostenveld, "Nonparametric statistical testing of EEG- and MEG-data," *Journal of Neuroscience Methods*, vol. 164, no. 1, pp. 177-190, 2007. doi: 10.1016/j.jneumeth.2007.03.024

[24] H. Wang, B. Li, Y. Lu, K. Han, H. Sheng, J. Zhou, Y. Qi, X. Wang, Z. Huang, L. Song, and Y. Hua, "Real-time threshold determination of auditory brainstem responses by cross-correlation analysis," *iScience*, vol. 24, no. 11, p. 103285, 2021. doi: 10.1016/j.isci.2021.103285

[25] S. Bogaerts, J. D. Clements, J. M. Sullivan, and S. Oleskevich, "Automated threshold detection for auditory brainstem responses: comparison with visual estimation in a stem cell transplantation study," *BMC Neuroscience*, vol. 10, p. 104, 2009. doi: 10.1186/1471-2202-10-104

[26] D. Thalmeier, G. Miller, E. Schneltzer, J. Hirst, M. Sharma, and T. Muller, "Objective hearing threshold identification from auditory brainstem response measurements using supervised and self-supervised approaches," *BMC Neuroscience*, vol. 23, 2022. doi: 10.1186/s12868-022-00758-0

[27] A. Schilling, R. Gerum, P. Krauss, C. Metzner, K. Tziridis, and H. Schulze, "Objective Estimation of Sensory Thresholds Based on Neurophysiological Parameters," *Frontiers in Neuroscience*, vol. 13, p. 481, 2019. doi: 10.3389/fnins.2019.00481

[28] K. Suthakar and M. C. Liberman, "A simple algorithm for objective threshold determination of auditory brainstem responses," *Hearing Research*, vol. 381, p. 107782, 2019. doi: 10.1016/j.heares.2019.107782

[29] D. Alpsan, M. Saffet, and O. Binici, "Determining hearing threshold from Brain Stem Evoked Potentials: optimising a neural network to improve classification performance," *Engineering in Medicine and Biology Magazine*, vol. 13, pp. 465-471, 1994. doi: 10.1109/51.310986

[30] N. Acır, Ö. Özdamar, and C. Güzeliş, "Automatic classification of auditory brainstem responses using SVM-based feature selection algorithm for threshold detection," *Engineering Applications of Artificial Intelligence*, vol. 19, no. 2, pp. 209-218, 2006. doi: 10.1016/j.engappai.2005.08.004

[31] E. Berninger, Å. Olofsson, and A. Leijon, "Analysis of click-evoked auditory brainstem responses using time domain cross-correlations between interleaved responses," *Ear and Hearing*, vol. 35, no. 3, pp. 318-329, 2014. doi: 10.1097/01.aud.0000441035.40169.f2

[32] B. K. Cone-Wesson, K. G. Hill, and G.-B. Liu, "Auditory brainstem response in tammar wallaby (*Macropus eugenii*)," *Hearing Research*, vol. 105, no. 1-2, pp. 119-129, 1997. doi: 10.1016/S0378-5955(96)00199-2

[33] A. Dobrowolski et al., "Classification of auditory brainstem response using wavelet decomposition and SVM network," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 2, pp. 427-436, 2016. doi: 10.1016/j.bbe.2016.01.003

[34] R. Zhang, G. McAllister, B. Scotney, S. McClean, and G. Houston, "Coupling Wavelet Transform with Bayesian Network to Classify Auditory Brainstem Responses," in *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, Shanghai, China, 2005, pp. 1-4. doi: 10.1109/IEMBS.2005.1616263

[35] P. J. McCullagh, H. Wang, H. Zheng, G. Lightbody, and H. G. McAllister, "A Comparison of Supervised Classification Methods for Auditory Brainstem Response Determination," in *MedInfo 2007*, 2007, pp. 1289-1293. doi: 10.3233/978-1-58603-774-1-1289

[36] E. Vannier, O. Adam, and J.-F. Motsch, "Objective detection of brainstem auditory evoked potentials with a priori information from higher presentation levels," *Artificial Intelligence in Medicine*, vol. 25, no. 3, pp. 283-301, 2002. doi: 10.1016/S0933-3657(02)00029-5

[37] M. P. Paulraj, K. Subramaniam, S. B. Yaccob, A. H. B. Adom, and C. R. Hema, "A machine learning approach for distinguishing hearing perception level using auditory evoked potentials," in *2014 IEEE Conference on Biomedical Engineering and Sciences (IECBES)*, Kuala Lumpur, Malaysia, 2014, pp. 991-996. doi: 10.1109/IECBES.2014.7047661

[38] R. Davey, P. McCullagh, G. Lightbody, and G. McAllister, "Auditory brainstem response classification: a hybrid model using time and frequency features". *Artif Intell Med*, May;40(1):1-14. doi: 10.1016/j.artmed.2006.07.001

[39] E. Stürzebecher, M. Cebulla, and C. Elberling, "Automated Auditory Response Detection: Statistical Problems with Repeated Testing," *International Journal of Audiology*, vol. 44, no. 2, pp. 110-117, 2005. doi: 10.1080/14992020400029228

[40] E. Stürzebecher and M. Cebulla, "Automated Auditory Response Detection: Improvement of the Statistical Test Strategy," *International Journal of Audiology*, vol. 52, no. 12, pp. 861-864, 2013. doi: 10.3109/14992027.2013.822995

[41] M. Cebulla and E. Stürzebecher, "Automated auditory response detection: Further improvement of the statistical test strategy by using progressive test steps of iteration," *International Journal of Audiology*, vol. 54, no. 8, pp. 568-572, Apr. 2015. doi: 10.3109/14992027.2015.1017659

[42] M.A. Chesnaye, S.L. Bell, J.M. Harte, and D.M. Simpson, "A group sequential test for ABR detection," *International Journal of Audiology*, vol. 58, no. 10, pp. 618-627, Oct. 2019. doi: 10.1080/14992027.2019.1625486

[43] M.A. Chesnaye, S.L. Bell, J.M. Harte, and D.M. Simpson, "The Convolutional Group Sequential Test; Reducing Test Time for Evoked Response Detection," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 3, March 2020. doi: 10.1109/TBME.2019.2919696

[44] T. Zanotelli, F. Antunes, D.M. Simpson, E.M.A.M. Mendesa, and L.B. Felix, "Faster automatic ASSR detection using sequential tests," *International Journal of Audiology*, vol. 59, no. 8, pp. 631-639, 2020. doi: 10.1080/14992027.2020.1728402

[45] Ö. Özdamar, R.E. Delgado, R.E. Eilers, and J.E. Widen, "Computer Methods for On-Line Hearing Testing with Auditory Brain Stem Responses," *Ear and Hearing*, vol. 11, no. 6, pp. 417-429, 1990. doi: 10.1097/00003446-199012000-00003

[46] P. Armitage, C.K. McPherson, and B.C. Rowe, "Repeated Significance Tests on Accumulating Data," *Journal of the Royal Statistical Society. Series A (General)*, vol. 132, no. 2, pp. 235-244, 1969. doi: 10.2307/2343787

[47] C.E. Rasmussen and K.I. Williams, *Gaussian Processes for Machine Learning*, Massachusetts Institute of Technology, Nov. 2006. ISBN: 026218253X

[48] I. Silva and M. Epstein, "Estimating loudness growth from tone-burst evoked responses," *J Acoust Soc Am*, vol. 127, issue 6, Jun 2010. doi: 10.1121/1.3397457

[49] R.B. Gramacy, *Surrogates: Gaussian Process Modeling, Design and Optimization for the Applied Sciences*. Chapman & Hall, Dec 2021. ISBN: 9781032242552

[50] A.C. Rencher, *Methods of Multivariate Analysis*, 2nd ed. New York: John Wiley and Sons, Inc., 2001. ISBN: 0-471-41889-7

[51] R.C. Littell and J.L. Folks, "Asymptotic Optimality of Fishers Method of Combining Independent Tests," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 802–806, Dec 1971. doi: 10.2307/2284230

[52] A.G. Wilson, "Covariance Kernels for Fast Automatic Pattern Discovery and Extrapolation with Gaussian Processes," PhD dissertation, Trinity College, University of Cambridge, 2014.

[53] N. Cressie, "The origins of kriging," *Mathematical Geology*, vol. 22, pp. 239–252, 1990. doi: 10.1007/BF00889887

[54] J. Gonzalvez, E. Lezmi, T. Roncalli and J. Xu, "Financial Applications of Gaussian Processes and Bayesian Optimization," *SSRN Electronic Journal*, Apr 2019. Corpus ID: 159451782. doi: 10.2139/ssrn.3344332

[55] V. Rajpaul, S. Aigrain, M.A. Osborne, S. Reece and S. Roberts, "A Gaussian process framework for modelling stellar activity signals in radial velocity data," *Monthly Notices of the Royal Astronomical Society*, vol. 452, issue 3, pp. 2269–2291, Sep 2015. doi: 10.1093/mnras/stv1428

[56] A. Arjas, A. Hauptmann and M.J. Sillanpää, "Estimation of dynamic SNP-heritability with Bayesian Gaussian process models," *Bioinformatics*, vol. 36, issue 12, pp. 3795–3802, June 2020. doi: 10.1093/bioinformatics/btaa199

[57] J. Xu and K. Suzuki, "Massive-training support vector regression and Gaussian process for false-positive reduction in computer-aided detection of polyps in CT colonography," *Med Phys*, vol. 38, issue 4, pp. 1888–902, Apr 2011. doi: 10.1118/1.3562898

[58] N. Best, S. Richardson and A. Thomson, "A comparison of Bayesian spatial models for disease mapping," *Statistical Methods in Medical Research*, vol. 14, issue 1, pp. 35-59, Feb. 2005. doi: 10.1191/0962280205sm388oa

[59] M. Strano, "A technique for FEM optimization under reliability constraint of process variables in sheet metal forming," *International Journal of Material Forming*, vol. 1, issue 1, pp. 13–20, Mar 2008. doi: 10.1007/s12289-008-0001-8

[60] J.Schlittenlacher, R.E. Turner and B.C.J. Moore, "Audiogram estimation using Bayesian active learning," *J Acoust Soc Am*, vol. 144, issue 1, Jul 2018. doi: 10.1121/1.5047436