AN ACTIVE LEARNING PROCEDURE FOR THE INTERAURAL TIME DIFFERENCE DISCRIMINATION THRESHOLD

Andrea Gulli, Federico Fontana
Department of Mathematics, Computer Science and Physics
Università di Udine, Udine, Italy
gulli.andrea@spes.uniud.it

Stefania Serafin
Department of Architecture, Design, and Media Technology
Aalborg University, Copenhagen, Denmark
sts@create.aau.dk

Michele Geronazzo
Department of Engineering and Management
Università di Padova, Padua, Italy
michele.geronazzo@unipd.it

ABSTRACT

Measuring the auditory lateralization elicited by interaural time difference (ITD) cues involves the estimation of a psychometric function (PF). The shape of this function usually follows from the analysis of the subjective data and models the probability of correctly localizing the angular position of a sound source. The present study describes and evaluates a procedure for progressively fitting a PF, using Gaussian process classification of the subjective responses produced during a binary decision experiment. The process refines adaptively an approximated PF, following Bayesian inference. At each trial, it suggests the most informative auditory stimulus for function refinement according to Bayesian active learning by disagreement (BALD) mutual information. In this paper, the procedure was modified to accommodate two-alternative forced choice (2AFC) experimental methods and then was compared with a standard adaptive "three-down, one-up" staircase procedure. Our process approximates the average threshold ITD 79.4% correct level of lateralization with a mean accuracy increase of 8.9% over the Weibull function fitted on the data of the same test. The final accuracy for the Just Noticeable Difference (JND) in ITD is achieved with only 37.6% of the trials needed by a standard lateralization test.

1. INTRODUCTION

The ability to localize sound sources is of considerable importance for humans and animals; it determines the direction of objects to be sought or avoided and the appropriate direction to direct visual attention. Although auditory localization may rely on the sound arriving at one ear, the most reliable localization cues depend on the acoustic waves arriving at both ears [1]. The difference between the two paths from a sound source to the ears creates an interaural time difference (ITD). In parallel, an interaural level difference (ILD) occurs due to the head shadow on the contralateral ear. In humans, the cue that enables sound localization most accurately (up to 1 degree in azimuth) is the ITD [2]. Experiments capable of isolating ITD used pairs of "on the ear" stimulators, namely headphones. Early headphone-based tests reported ITD detection thresholds at microsecond scale, that is, orders of magnitude smaller than all other sensory modalities were able to detect [3].

When headphones are worn, the sound source image is localized inside the head. The term "lateralization" was then adopted to describe the apparent sound source position inside the head. On the other hand, headphones allow for precise control of interaural differences and do not generate room echoes. Therefore, lateralization were preferred to localization tests when studying sound source perception inside the laboratory [4]. Since the 1950s, accurate studies have been systematically conducted to determine ITD thresholds [5]. The lowest were reported to be close to 10 μs. However, the participants' hearing and training level necessary to achieve those thresholds have become clear only recently, along with the stimuli and measurement technique required for measuring them [6]. Specifically, the stimulus that produced the lowest ITD threshold was Gaussian noise, bandpass filtered from 20 to 1400 Hz and presented at a sound pressure level of 70 dB. The most accurate method was a two-interval procedure with an interstimulus interval of 50 ms. The mean ITD threshold in this condition at the 75% corrected level was 6.9 μs for trained listeners, and 18.1 μs for untrained listeners. However, other studies report higher ITD values as *normal* hearing thresholds, i.e., with mean equal to 263 μs and standard deviation equal to 112 μs [7].

We present an accelerated procedure for reliably determining individual lateralization thresholds, and compare it to standard approaches to subjective ITD measurement. We will focus on untrained participants with the goal of significantly shortening the test sessions. This feature would be desirable especially when specific groups of users are targeted, such as young patients whose binaural acuity needs to be tested. Moreover, the same procedure can quickly calibrate and individualize immersive audio technologies for the most diverse applications and virtual environments [8].

1.1. Psychometric function estimation

A psychometric function (PF) maps the subjective performance during a perceptual task against a stimulus magnitude, such as brightness or other intensity levels. Performance is measured as the percentage of correct responses, or responses where the participant was able to detect the stimulus. Ideally, a PF is estimated at informative sample points on a continuous scale. The level set estimation (LSE) problem consists of identifying the regions where an initially unknown PF $\mathrm{f}(x)$ lies above or below a particular threshold ϑ. In general, a level set S is the set on which f exceeds some critical value (e.g., $S = x : \mathrm{f}(x) > \vartheta$). Efficient LSE is an active learning problem [9], involving techniques that use surrogate models to perform active sampling [10]. The active learning configuration consists of the definition of an acquisition function that classifies the data points to be labeled according to the current state of the model and a hand-designed information measure to be maximized [11]. In Bayesian active-learning (BAL), the basic idea is to define a statistical model and then tune its parameters in due data

collection. Typically, the model is initialized with a weakly informative prior distribution which expresses the uncertainty about these parameters before the start of the experiment. Then, recorded data provide likelihood terms to be combined with the prior in a posterior distribution, reflecting the beliefs about the parameters from the data collected so far. The stimulus for every subsequent trial is selected so as to maximize some utility measure that is integrated with the current posterior. One of the first BAL procedures in psychophysics that used Bayesian principles for both modeling the response and choosing the parameters for the next trial [12] was designed to classify a subject into one of nine audiometric groups, and was then validated with numerical simulations. The stimulus for the subsequent trial was chosen to maximize the mutual information between the current and the following unknown estimate by selecting it with the least expected entropy [13]. Selection was made by computing the posterior probabilities across all candidate stimuli for the next trial.

A general BAL procedure for classification and preference tasks that uses Gaussian Processes (GP) [14] for estimating a subjective response is called Bayesian Active Learning by Disagreement (BALD) [15]. These are the approximation technique and the acquisition function employed in this study. GP-based Bayesian inference has been recently employed in machine learning applications across various disciplines [16], and specifically in audiology [17, 18]. GPs in fact incorporate prior hypotheses about the mean, the smoothness between class boundaries, and the covariance between data points. BALD active learning with GP classification (GPC) has already been used in auditory applications, e.g. for optimal setting of a hearing aid [19], and for determining audiograms [20, 21], equal-loudness contours [22], and psychometric functions [23]. However, it has never been employed in the measurement of ITD thresholds.

This paper presents: Sec. 2 the mathematical background of GP and BALD classification; Sec. 3 the characteristics of the specific model; Sec. 4 the test determining individual lateralization thresholds with BALD; Sec. 5 the results, and Sec. 6 their discussion. Sec. 7 concludes the paper. As we will see from our results, GPC with active learning is a valid approximation for the PF, with a RMSE computed on the whole test set which is smaller than 10% concerning the conventional Weibull fitting [24]; it achieves similar performances as the default procedure (mean error equal to 5.1 μs) by requiring only 37.6% of the trials otherwise needed by the standard procedure.

2. THEORETICAL BAKGROUND

2.1. Gaussian Processes Classification

Let $\mathrm{f} : \mathbb{R} \rightarrow \mathbb{R}$ be a latent function on an arbitrary input space $\mathbf{X}$. A GP is a convenient technique for encoding prior knowledge about f that can be later updated via Bayesian inference in light of the observed data. A GP is a collection of random variables, any finite subset of which jointly forms a Gaussian distribution. Therefore, a GP is a particular case of a stochastic process. Like the multivariate Gaussian distribution, a GP is completely specified by its first two moments: a mean function $\mu(\mathbf{x})$ and a positive semidefinite covariance function $K(\mathbf{x}, \mathbf{x}')$. The mean function expresses the central tendency of the latent function, while the covariance function accounts for its correlation structure. Given μ and K, the latent function f can be endowed with a GP prior distribution

$$p(\mathrm{f}) = \mathcal{GP}\left(\mu(\mathbf{x}), K(\mathbf{x}, \mathbf{x}')\right). \tag{1}$$

Given a GP prior on f and some observations over the input space, a prediction can be performed about the behavior of f for unobserved inputs using Bayesian inference, computed by Bayes' rule:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}.$$

According to Bayes' theorem, the joint posterior of the latent function at training and test inputs given the training observations is

$$p(\mathbf{f}, \mathrm{f}_*|\mathbf{X}, \mathbf{y}, x_*) = \frac{p(\mathbf{f}, \mathrm{f}_*|\mathbf{X}, x_*)p(\mathbf{y}|\mathbf{f})}{p(\mathbf{y}|\mathbf{X})}, \tag{2}$$

where $\mathrm{f}(x_*) = \mathrm{f}_*$, and $\mathbf{f} = \mathrm{f}(\mathbf{X})$. The predictive posterior distribution can be determined by marginalizing out the training set latent variables and substituting in (2),

$$\begin{aligned} p(\mathrm{f}_*|\mathbf{X}, \mathbf{y}, x_*) &= \int p(\mathbf{f}, \mathrm{f}_*|\mathbf{X}, \mathbf{y}, x_*)d\mathbf{f} \\ &= \frac{1}{p(\mathbf{y}|\mathbf{X})} \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}, \mathrm{f}_*|\mathbf{X}, x_*)d\mathbf{f}, \end{aligned} \tag{3}$$

and by definition of the GP, the joint probability $p(\mathbf{f}, \mathrm{f}_*|\mathbf{X}, x_*)$ is a multivariate Gaussian. With the posterior, we can compute a probabilistic prediction of the latent function f at the new input locations $\mathbf{X}_*$, taking into consideration the previously observed samples $(\mathbf{y}, \mathbf{X})$. The posterior mean and the posterior covariance on f provide information about the updated beliefs and the remaining uncertainty about the latent function. The likelihood $p(\mathbf{y}|\mathbf{f})$ describes the relationship between the latent function values $\mathbf{f}$ and the observations $\mathbf{y}$ at the training inputs $\mathbf{X}$.

The focus of this study is one-dimensional binary classification, where observed outputs can only assume two values: 1 (success) or 0 (failure). The latent function f is not directly observed but is instead a hidden function, where larger values of f generate higher probabilities of success. To obtain the probabilistic distribution $p(y = 1|\mathrm{f})$, f is "squashed" using a monotonically increasing sigmoid function Φ to the range [0,1]. For a binary observation y_i associated with an input $x_i \in \mathbf{X}$,

$$p(y_i = 1|\mathrm{f}) = \Phi(\mathrm{f}_i) = \Phi\left(\mathrm{f}(x_i)\right). \tag{4}$$

One largely used and convenient choice of Φ to deal with binary classification problems is the inverse-logit, also known as Bernoulli-logistic function likelihood, given by

$$\Phi(\mathrm{f}_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mathrm{f}_i} e^{-\frac{t^2}{2}} dt, \tag{5}$$

and assuming that the labels $\mathbf{y} = (y_1, \ldots, y_N)$ of the N training data points are conditional independent if (latent) $\mathbf{f}$ are known,

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} p(y_i|\mathrm{f}(x_i)). \tag{6}$$

The predictive posterior distribution in (3) with (6) now becomes

$$p(\mathrm{f}_*|\mathbf{X}, \mathbf{y}, x_*) = \frac{1}{p(\mathbf{y}|\mathbf{X})} \int \prod_{i=1}^{N} \Phi(\mathrm{f}_i)p(\mathbf{f}, \mathrm{f}_*|\mathbf{X}, x_*)d\mathbf{f}. \tag{7}$$

From (7), the probabilistic prediction of class identity for a test observation y_* can be computed as:

$$\begin{aligned} p(y_* = 1|\mathbf{X}, \mathbf{y}, x_*) &= \int p(y_* = 1|\mathrm{f}_*)p(\mathrm{f}_*|\mathbf{X}, \mathbf{y}, x_*)d\mathrm{f}_* \\ &= \int \Phi(\mathrm{f}_*)p(\mathrm{f}_*|\mathbf{X}, \mathbf{y}, x_*)d\mathrm{f}_*. \end{aligned} \tag{8}$$

The non-Gaussian likelihood of (5) for the classification framework given by (4) makes the integrals in (7) and (8) analytically intractable. Dropping the conditioning on the training and test data points for ease of notation and applying the reverse chain rule, we have

$$p(\mathbf{f}, \mathrm{f}_*|\mathbf{y}) = p(\mathrm{f}_*|\mathbf{f})p(\mathbf{f}|\mathbf{y}). \tag{9}$$

The first term of (9) can be computed by applying the multivariate Gaussian conditional rule to the GP prior, while the second term can be approximated with variational inference [25]. A multivariate Gaussian variational distribution $q(\mathbf{f})$ approximating the posterior $p(\mathbf{f}|\mathbf{y})$ is found through the minimization of the Kullback–Leibler divergence (KL divergence) [26] $\mathrm{KL}(q(\mathbf{f})||p(\mathbf{f}|\mathbf{y}))$. Since this similarity measure is also intractable, the variational evidence lower buond (ELBO) is used as a proxy for the KL divergence minimization:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} = \; & \mathbb{E}_{p(y,\mathbf{x})}\left[\mathbb{E}_{p(\mathrm{f}|\mathbf{u},\mathbf{x})q(\mathbf{u})}\left[\log p(y\,|\,\mathrm{f})\right]\right] \\ & - \; \mathrm{KL}\left[q(\mathbf{u})\|p(\mathbf{u})\right] \\ \approx \; & \sum_{i=1}^{N} \mathbb{E}_{q(\mathrm{f}_i)}\left[\log p(y_i\,|\,\mathrm{f}_i)\right] - \; \mathrm{KL}\left[q(\mathbf{u})\|p(\mathbf{u})\right], \end{aligned} \tag{10}$$

where N is the number of data points, $q(\mathbf{u})$ is the Gaussian variational distribution computed at the inducing function values $\mathbf{u}$, $q(\mathrm{f}_i)$ is the marginal of $p(\mathrm{f}_i|\mathbf{u}, x_i)q(\mathbf{u})$, $p(\mathbf{u})$ is the GP prior distribution for the inducing function values. The ELBO is the lower bound of the log marginal likelihood $\log(p(\mathbf{y}))$, also called model evidence, and it is an expression containing all the parameters defining the GP prior and the variational distribution; thus, gradient descent can be used to maximize the ELBO concerning the model parameters to find concrete values for those parameters. In practice, the negation of (10) will be used as the "loss" function to determine the "hyperparameters" of the GP prior distribution.

2.2. Bayesian Active Learning by Disagreement

The fundamental principle of active learning requires that a model actively selects input queries $x_i \in \mathbf{X}$ and observes the system's response y_i, rather than passively collecting (x_i, y_i) pairs. The goal of information theoretic active learning is to reduce the number of possible hypotheses in the fastest way, i.e., to minimize the uncertainty about the parameters using Shannon's entropy [27]. The objective is to seek the data point $\mathbf{x}$ that maximizes the decrease in expected posterior entropy [15]:

$$\arg\max_{\mathbf{x}} \mathrm{H}[\boldsymbol{\theta}|\mathcal{D}] - \mathbb{E}_{\,y\sim p(y|\mathbf{x},\mathcal{D})}[\mathrm{H}[\boldsymbol{\theta}|\mathcal{D}]],$$

or, equivalently, the point maximizing the conditional mutual information between the unknown output and the parameters $\boldsymbol{\theta}$, given a training dataset $\mathcal{D}$:

$$\arg\max_{\mathbf{x}} \mathrm{H}[y|\mathbf{x},\mathcal{D}] - \mathbb{E}_{\,\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}[\mathrm{H}[y|\mathbf{x},\boldsymbol{\theta}]].$$

BALD searches the $\mathbf{x}$ for which the model is marginally most uncertain about y, holding confident individual settings of the model parameters.

The BALD algorithm for GPC consists of two steps. First, it applies an approximate inference algorithm for GPCs to obtain the posterior predictive mean $\mu_{\mathbf{x},\mathcal{D}}$ and variance $\sigma^2_{\mathbf{x},\mathcal{D}}$ for each point of interest $\mathbf{x}$. Then, it selects a query $\mathbf{x}$ that maximizes the mutual information. The first term can be expressed in terms of the binary entropy function h:

$$\begin{aligned} \mathrm{H}[y|\mathbf{x},\mathcal{D}] &\approx \mathrm{h}\left(\int \Phi(\mathrm{f}_\mathbf{x})\mathcal{N}(\mathrm{f}_\mathbf{x}|\mu_{\mathbf{x},\mathcal{D}}, \sigma^2_{\mathbf{x},\mathcal{D}})d\mathrm{f}_\mathbf{x}\right) \\ &= \mathrm{h}\left(\left(\Phi\left(\frac{\mu_{\mathbf{x},\mathcal{D}}}{\sqrt{\sigma^2_{\mathbf{x},\mathcal{D}}+1}}\right)\right)\right), \end{aligned} \tag{11}$$

with $\mathrm{h}(p) = -p\log p - (1-p)\log(1-p)(p)$. The second term $\mathbb{E}_{\,\mathrm{f}\sim p(\mathrm{f}|\mathcal{D})}[\mathrm{H}[y|\mathbf{x},\boldsymbol{\theta}]]$ can be approximated to

$$\mathbb{E}_{\,\mathrm{f}\sim p(\mathrm{f}|\mathcal{D})}[\mathrm{H}[y|\mathbf{x},\boldsymbol{\theta}]] \approx \frac{C}{\sqrt{\sigma^2_{\mathbf{x},\mathcal{D}}+C^2}} \exp\left(-\frac{\mu^2_{\mathbf{x},\mathcal{D}}}{2\left(\sigma^2_{\mathbf{x},\mathcal{D}}+C^2\right)}\right),$$

where $C = \sqrt{\pi \ln 2/2}$.

3. GPC FOR 2AFC

Finding the most accurate way to approximate the PF describing lateralization ability is the main objective of this work. A particularly important value in two alternative forced choice (2AFC) tests is the point at which the PF assumes a certain percentage, typically 70.7%, 75%, 76%, or 79.4% [28], with which the ITD threshold is associated. In these psychometric tests, the probability of assigning the correct classification label to a specific stimulus cannot be less than 50%, and the maximum probability must consider a percentage of errors given by a lapse rate close to zero [29].For these reasons the Gaussian-modeled latent function is squashed into the interval [0.5, 1]. The mean and covariance functions of the GP prior given by (1) are respectively set to a constant function $\mu(x) = \mu$, and the radial basis function

$$K(x, x') = \exp\left(-\frac{1}{2}(x-x')^\top \theta^{-2}(x-x')\right), \tag{12}$$

where θ is a length scale parameter. The hyperparameters' vector $\boldsymbol{\theta} = (\mu, \theta)$ is determined during the training, i.e., during the minimization of the negative log likelihood given by the negation of the variational ELBO in (10). The variational ELBO is modeled to approximate the likelihood once it is scaled to the restricted probability values. This likelihood is computed as

$$\Phi\left(C'\right) = \frac{1}{4}\left(1 + \mathrm{erf}\left(\frac{C'}{\sqrt{2}}\right)\right) + \frac{1}{2}, \tag{13}$$

where $C' = \mu_{\mathbf{x},\mathcal{D}}/\sqrt{\sigma^2_{\mathbf{x},\mathcal{D}}+1}$, $\mu_{\mathbf{x},\mathcal{D}}$ and $\sigma^2_{\mathbf{x},\mathcal{D}}$ are respectively the mean and the variance of the Gaussian-modeled latent function, and erf is the error function:

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z \exp -t^2 dt. \tag{14}$$

To take this scaling into account while computing the acquisition function of the BALD procedure, we rescale the likelihood between 0 and 1 at the Shannon's entropy input. This way, a maximum entropy starting around the 75% probability point is obtained.

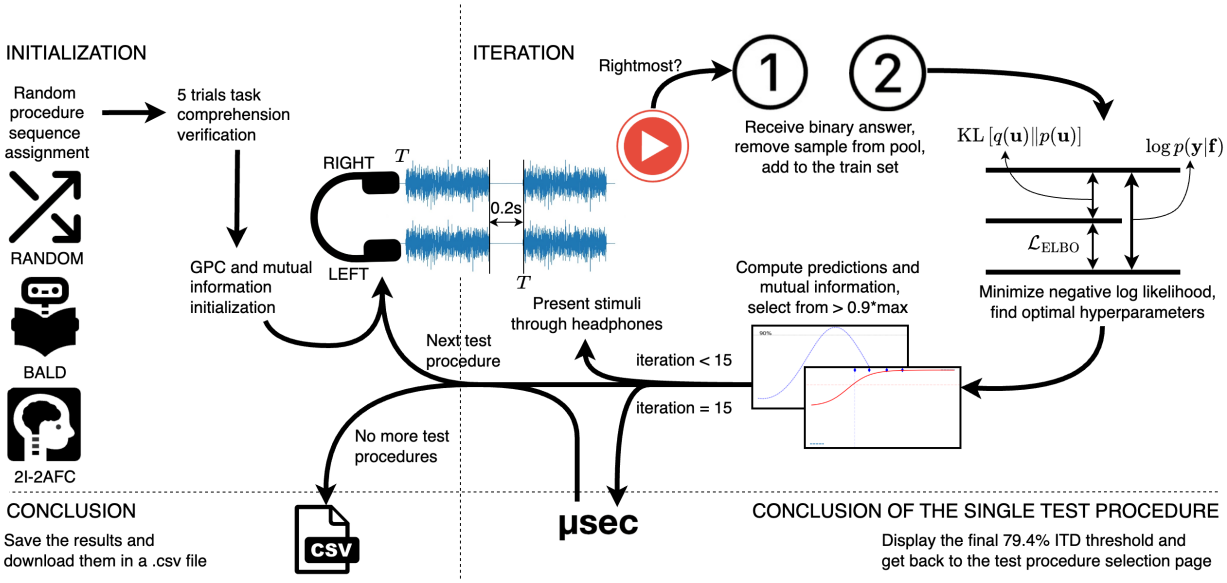

Figure 1: *Test protocol and experimental flow.*

4. THE EXPERIMENT ON ITD DISCRIMINATION THRESHOLDS

4.1. Participants

Seventeen young adults (5 male and 12 female, mean age: 31.82 ± 6.38 years) reporting no hearing deficiencies participated in the experiment. Twelve had audiometric thresholds equal to or less than 20 dB hearing level (HL) at octave-spaced frequencies from 125 to 8000 Hz. For those five participants without audiometry, the answers to Sanders' questionnaire [1] reported no hearing difficulties. One participant was excluded from the experiment since reporting to be unable to concentrate sufficiently during the test. All participants reported no prior experience with a binaural hearing test.

4.2. Acoustic Stimulus and Apparatus

Narrowband noise was synthesized at 10 MHz sampling rate, using the Python TorchAudio software package [30]. Band-pass filtering was performed in the frequency domain so as to limit the noise frequency band to the [20–1400] Hz range [6]. The amplitude level was calibrated to 70 ± 1 dB SPL using an NTi Audio XL2 sound level meter. After temporal gating, a short noise burst lasting 0.5 s was created having a 50 ms squared-cosine onset and offset.

A burst sequence was formed by intertwining identical noise bursts with silence lasting 0.2 s. The stimulus was formed by pairing a delayed version of this sequence and a new version of the same sequence, obtained by zero-padding the original until reaching the same length as the former. The delay could be varied so as to define a desired ITD $= T_R - T_L = T - (-T) = 2T$:

$$\begin{aligned} \text{RIGHT} \quad & \overbrace{0,\ldots,0}^{T}, \text{burst}, \overbrace{0,\ldots,\ldots,0}^{0.2\text{ s}}, \text{burst}, \overbrace{0,\ldots,0}^{\text{padding}} \\ \text{LEFT} \quad & \text{burst}, \underbrace{0,\ldots,0}_{\text{padding}}, \underbrace{0,\ldots,\ldots,0}_{0.2\text{ s}}, \underbrace{0,\ldots,0}_{T}, \text{burst} \end{aligned} . \tag{15}$$

The stimulus was presented on two audio channels through a pair of AKG K240 MKII semi-open headphones, whose frequency response was flattened using the AutoEQ software[2]. Sounds were reproduced by a 13" MacBook Pro M2 laptop computer after sampling them down to 96 kHz, through interpolation with a sync function windowed by a Hann window. This sampling rate was the highest available in the laptop's audio interface, in practice limiting the lowest ITD to 10 μs. A GUI enabling attendance to the task was realized in HTML, CSS, and JavaScript programming languages as a custom Flask web application. The test took place in a room with background noise equal to 20 ± 2.5 dB SPL.

4.3. Task and Experimental Protocol

The experimental protocol is illustrated in Fig. 1. During the test, each participant was sitting in front of the laptop computer running the GUI. At each trial, the task consisted of listening to two subsequent and randomly balanced stimuli, and then choosing which sound was the rightmost, by selecting it with the mouse on the computer screen. At the beginning of each session, five pilot trials were presented having ITD levels equal to 240, 200, 160, 120, and 80 μs. Correct guesses in all such trials were necessary for the measurements to start in correspondence with the sixth trial. A session lasted approximately 10 minutes.

The protocol was designed for determining the PF 79.4% threshold, describing the subjective lateralization performance as a function of T defined in (15). The target (i.e., rightmost) stimulus

[1] https://www.aooi.it/contents/attachment/c4/ref121.pdf, a validated questionnaire to evaluate the actual level of communication in various situations, e.g., at home or in a social environment, accessed Feb 28, 2023.

[2] https://githubq.com/jaakkopasanen/AutoEq, accessed Feb 28, 2023.

was lateralized twice as much as a nominal ITD, and the reference source (i.e., leftmost) stimulus was instead lateralized with an opposite ITD. Hence, an ITD threshold equal to $2T$ means that a participant discriminated the target ITD by $T\mu$s from the reference ITD of $-T\mu$s. The presentation of symmetric ITDs minimizes hemispheric effects, and ensures that participants could not perform the task based on perceived changes in interaural coherence [28].

The protocol implemented three different procedures:

- adaptive two-interval 2AFC (2I-2AFC hereafter),
- GPC with BALD active learning (BALD hereafter),
- GPC with random acquisition function (RANDOM hereafter).

Accordingly, every session included three series of trials respectively implementing such procedures in a randomly balanced order. When GPC was used, thus enabling active learning and random selection of ITDs, the number of trials was empirically set to 15. We compared these three procedures to observe the efficacy of GPC in ITD threshold estimation and the contribution of the BALD algorithm. In the standard 2I-2AFC procedure, individual thresholds were computed using the adaptive track reversal technique or a Weibull fitting of the participants' answers. We consider this procedure as a reference since it is commonly used to determine JND thresholds for many types of stimuli, moreover it can be designed to converge on a desired percentage. Conversely, the GPC approximation was compared to a Weibull fitting of the data collected using the other two procedures. Reversals were included in the comparison even though they are known to have high bias and smaller precision [31], considering their difference from a Weibull fitting in the same manner as it happens for the upper limit of a reasonable estimate. The RANDOM procedure was implemented to test the GPC technique separately from active learning, i.e., to analyze the GPC on training points other than highly informative data points. Assuming a participant performs consistently across all tests, the RANDOM procedure will converge, but in a variable number of iterations dependent on the samples' random distribution.

4.3.1. Default 2I-2AFC

An adaptive "three-down, one-up" staircase procedure was chosen, i.e., T was decreased after three correct responses and increased after one wrong response. Theoretically this procedure estimates the 79.4% correct level on the PF [32]. The corresponding series of trials started with an ITD threshold that was above the threshold estimated during the pilot series, equal to 80 μs, i.e., presenting stimuli with $T = 40$ μs. T step-size was initially a factor of 2 and then reduced to 1.414 and 1.189 after the first and second "down-up reversal", respectively. This series terminated after six reversals at the smallest step size. T was varied by logarithmic steps [32].

Psychometric functions were estimated using a parametric fit of a Weibull function to all responses with a non-linear least squares optimization:

$$\Psi(x;\gamma,\lambda) = \gamma + \frac{(1-\gamma-\lambda)}{2}\left(1 + \text{erf}\left(\frac{x-\mu_\Psi}{\sqrt{2\cdot\sigma_\Psi^2}}\right)\right), \quad (16)$$

where γ is the guess rate (i.e., 0.5), λ is the miss rate in the range [0.01,0.05], and erf is the error function of (14).

4.3.2. GPC

The GPC started with a training set of 20 points: 10 points between 0 μs and 9 μs, all labeled as wrong answers, and 10 points between 91 μs and 100 μs, all labeled as correct answers. T corresponding to 79.4% of correct answers was found to allow the fitted curve to reach the closest value to that percentage up to 1 μs. In Fig. 2, the two plots respectively represent the predictions of the GPC and Weibull fit, on a test set consisting of 100 equally-spaced points between 1 and 100 μs. Both identify the 79.4% correct answers point with a difference equal to 8 μs.

The GPC training, i.e., the hyperparameters' vector $\boldsymbol{\theta}$ optimization, was performed on the normalized data (mean equal to 0 and standard deviation equal to 1) and furthermore constrained to search only positive values. In this regard, the Adam optimizer [33] was employed with a learning rate equal to 0.1 and a number of iterations set to 300. The GPC was implemented in GPyTorch [34], a software platform for scalable GP inference built on PyTorch.

4.3.3. BALD

The BALD acquisition function was computed at each step to determine the stimulus for the next trial iteratively. T of the binaural stimulus was randomly selected among all possible points in the pool, achieving mutual information levels higher than 90% of its maximum value in that set. The initial pool was made to correspond to the test set, i.e., the 100 points between 1 and 100 μs. Once a sample was labeled, it was removed from the pool so that each point was labeled once. After the first random selection, a new sample was chosen from that restricted subset of the pool being at least 5 μs far from every previously selected sample. If no sample satisfied this condition, a random choice was made from the samples achieving mutual information levels higher than 90% of its maximum value in the current pool's subset. The first T value was randomly drawn between 46 μs and 64 μs.

5. RESULTS

The number of trials during the 2I-2AFC procedure had a mean value across participants equal to 39.94 and a standard deviation

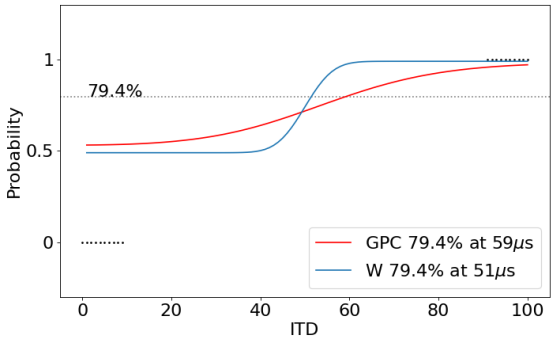

Figure 2: *Starting training data for the GPC (black dots) and predictions of both the GPC (red line) and the fitted Weibull function (W, blue line). The dotted horizontal line indicates the 79.4% level of correct answers.*

Table 1: *Individual 79.4% thresholds and means (μ) and standard deviations (σ) across participants for each technique. The "Rev" column lists T values found with the reversals procedure.*

#	@ 79.4% (μs)					
	2I-2AFC		BALD		RANDOM	
	W	Rev	W	GPC	W	GPC
1	10	14.1	6	1	48	34
2	57	67.2	54	54	56	43
3	25	33.6	23	19	31	22
4	14	11.9	16	10	1	10
5	22	33.6	34	26	2	17
6	39	40.0	57	41	16	13
7	30	33.6	35	23	16	8
8	39	47.6	38	32	57	39
9	46	67.2	48	53	52	52
10	23	28.3	26	22	25	23
11	40	47.6	38	32	54	44
12	10	14.1	19	17	20	15
13	58	67.3	52	55	47	30
14	21	23.8	23	18	20	15
15	39	47.6	51	36	57	66
16	45	47.6	55	49	38	22
17	31	33.6	29	23	39	39
μ	32.3	38.8	35.5	30.1	34.1	28.9
σ	14.4	17.4	15.0	15.5	18.7	15.9

equal to 5.70. The correct percentage had a mean equal to 80.20% and standard deviation equal to 3.85%. The repeated measures ANOVA test conducted on the number of trials asserted that the difference between the averages is big enough to be statistically significant ($p < 0.001$), and the magnitude of the difference between the averages is large (effect size $\eta > 0.9$). Table 1 shows the 79.4% ITDs for every test procedure, computed with the two respective aforementioned techniques. Mauchly's test of sphericity indicated that the assumption of sphericity had been violated, both for the individual T values ($\chi^2(14) = 74.11$, $p < 0.001$), and their logarithms ($\chi^2(14) = 92.848$, $p < 0.001$), therefore a Greenhouse-Geisser correlation was used (respectively, $\epsilon = 0.321$, and $\epsilon = 0.338$). The repeated measures ANOVA test did not reveal a significant main effect of either the fitting technique, or the testing procedure ($p > 0.1$). The means and standard deviations across participants of the differences of the 79.4% threshold points are shown in Table 2, along with their absolute values, between approximation techniques and procedures. The former reveals the tendency of the approximation to return an optimistic rather than pessimistic estimate of the ITD threshold, while the latter gives a measure of the divergence between the two approximations.

The comparison between the PF curve found with the Weibull function fitted to the reference test procedure data and the GPC, and the same Weibull fitting on the BALD and the random sampling procedures has been made using the root mean square error (RMSE). The lowest RMSE was found between the Weibull fittings in the reference test procedure and the BALD procedure ($\mu = 6.39\%, \sigma = 2.44\%$), followed by the GPC in the same test procedure ($\mu = 7.10\%, \sigma = 2.73\%$). Both the approximation techniques in the random procedure have a mean RMSE above 10% (Weibull: $\mu = 10.81\%, \sigma = 5.20\%$, GPC: $\mu = 11.10\%, \sigma = 4.65\%$). Figure 3 displays the data of a single individual collected in a complete test session.

Table 2: *Means (μ) and standard deviations (σ) of the differences (Δ) and the absolute values of the differences ($|\Delta|$) of the 79.4% T values found in the three procedures with the three techniques. Each approximation was compared with the one provided by the Weibull fitting on the same data, as well as with the approximation computed by that fitting on the reference 2I-2AFC test procedure ("W_{REF}").*

Procedure	Methods	Δ (μs)		$\lvert\Delta\rvert$ (μs)	
		μ	σ	μ	σ
2I-2AFC	Rev, W	6.5	5.1	6.7	4.8
BALD	GPC, W_{REF}	-2.2	5.1	5.1	2.4
	W, W_{REF}	3.2	6.6	5.6	4.7
	GPC, W	-5.5	5.3	6.4	4.1
RANDOM	GPC, W_{REF}	-3.4	15.4	12.1	10.1
	W, W_{REF}	1.8	15.2	12.4	9.1
	GPC, W	-5.1	9.3	9.0	5.7

6. DISCUSSION

The mean values displayed in Table 1 are in line with the thresholds found in literature [7]. Particularly in our experiment, in which stimuli with a sampling frequency below 100 KHz were reproduced, no ITDs smaller than 10 μs could be presented to the listeners. However, our participants were not trained for the specific task, and hence they were not expected to perceive ITD thresholds below this value. Conversely, it was demonstrated that humans can improve their lateralization skills after a dedicated training [35, 28].

The ITD is typically allowed to vary on a logarithmic scale in the lateralization literature, and data analysis is typically done using geometric means and standard deviation [36, 28]. However, alternative methodologies [35, 37] employ different test conditions in terms of the stimuli presented to the participant, and the use of logarithmic scaling allowed them to find one individual psychometric curve. Second, the goal of those experiments was to determine the smallest perceivable ITD, hence the use of the geometric mean and standard deviation to "zoom in" on the end scale of that measure. On the other hand, we want to determine the individual ITD thresholds with the same accuracy for any level of lateralization ability, particularly considering a possible application of the proposed procedure in audiology tests. Moreover, we expect that the same individual would perform similarly in each test because the stimulus and the task were kept unchanged during the whole session. Thus, data analysis is performed with an arithmetic mean and standard deviation.

The differences Δ in Table 2 reveal that the GPC with the BALD procedure is on average the closest 79.4% approximation to the reference 2I-2AFC procedure with the specific Weibull fitting (5.1 μs), providing a mean accuracy increase of 8.9% over the Weibull function fitted to the same data. This difference has the lowest standard deviation (2.4 μs). The GPC in the BALD procedure scores second for what concerns the signed difference (-2.2 μs), revealing an optimistic tendency; only the Weibull fitting in the random test procedure has a smaller one (1.8 μs), however the former has the lowest standard deviation while the latter has one among the highest, hence reflecting the prominent role of chance associated with that test procedure. The standard deviations of the differences Δ of the RANDOM procedure reflect the random number of iterations required by the GPC in that procedure to converge.

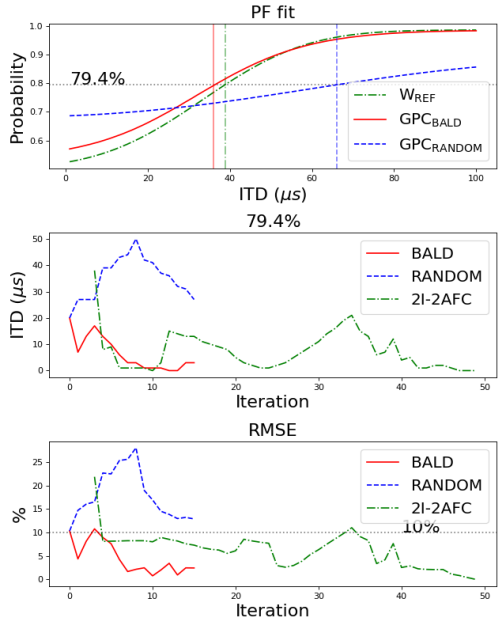

Figure 3: *Data of a single individual collected in a complete session. The top figure shows the GPC PF fittings and* $\mathrm{W_{REF}}$ *fitting. In the middle figure are displayed the absolute values of the differences of the 79.4% threshold points, while the bottom figure shows the RMSEs between the GPCs and the* $\mathrm{W_{REF}}$*, computed at every iteration step of the three procedures. The Weibull fitting in the 2I-2AFC procedure starts at the third iteration because the covariance of the parameters could not be estimated with less than three data points.*

Even if the analyses of variance did not reveal any statistically significant main difference, the BALD selection of samples is also valuable for the Weibull fitting since both the approximations used on that procedure's data have the smallest differences in absolute value and the least dispersed as well.

The RMSEs reported at the end of Sec. 5 confirm the suitability of GPC to approximate the PF in a 2AFC procedure and the ability of the BALD algorithm to identify optimal sampling; a further hint of this is that the standard deviations of the RANDOM procedure are the highest.

Future work may explore different directions. The proposed experimental protocol may be employed to test pre-trained individuals instead of novices and feed the GPC with a logarithmically transformed input to observe which fit gives the best results. Another direction may evaluate the BALD procedure without validated prior knowledge, such as for the hearing impaired, and check whether the starting training set needs to be modified or if other approximations are more suitable than GPC. The web application was chosen to share the test between different institutions and audiological research labs, allowing them to conduct the experiment remotely. In the current study, the application was used on a local server. The next step will include a thorough verification for deployment on a cloud platform.

7. CONCLUSIONS

In this study, a test for the fast determination of the Just Noticeable Difference in interaural time differences has been proposed and evaluated. A psychometric function was progressively fitted using Gaussian process classification of the subjective responses and Bayesian active learning by disagreement, aptly modified to accommodate a two-alternative forced choice experimental procedure. The results of its comparison with a standard adaptive "three-down, one-up" staircase procedure show that our process computes the closest approximation of the average threshold ITD 79.4% correct level of lateralization concerning the commonly used Weibull fitting on the reference test, with a mean accuracy increase of 8.9% over the Weibull function fitted on the data of the same test. The final accuracy was achieved with only 37.6% of the trials the standard adaptive staircase procedure needs.

The data and the web application are freely available at `https://zenodo.org/record/7808559` and `https://github.com/gullogullo/ITDtest`, respectively.

8. ACKNOWLEDGMENTS

We thank the Multisensory Experience Laboratory at the Aalborg University, Denmark, for making this research possible. Ali Adjorlu suggested design ideas for the GUI. The staff of the Complex Structure of Otorhinolaryngology and Audiology at the Institute for Maternal and Child Health IRCCS "Burlo Garofolo" in Trieste, Italy, participated in the experiment.

9. REFERENCES

[1] Lord Rayleigh, "Xii. on our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, 1907.

[2] Allen William Mills, "On the minimum audible angle," *The Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 237–246, 1958.

[3] Otto Klemm, "Investigations of the localization of sound stimuli iv: On the influence of binaural time difference on localization," *Arch. Ges. Psychol.*, vol. 40, pp. 117–145, 1920.

[4] Brian CJ Moore, *An introduction to the psychology of hearing*, Brill, 2012.

[5] Roy G Klumpp and Herman R Eady, "Some measurements of interaural time difference thresholds," *The Journal of the Acoustical Society of America*, vol. 28, no. 5, pp. 859–860, 1956.

[6] Sinthiya Thavam and Mathias Dietz, "Smallest perceivable interaural time differences," *The Journal of the Acoustical Society of America*, vol. 145, no. 1, pp. 458–468, 2019.

[7] Kubo Tetsushi, Tetsushi Sakashita, Makoto Kusuki, Kazushi Kyunai, Keita Ueno, Chie Hikawa, Tadashi Wada, Toshiyuki Shibata, Yoshiaki Nakai, and Takeshi, "Sound lateralization and speech discrimination in patients with sensorineural hearing loss," *Acta Oto-Laryngologica*, vol. 118, no. 543, pp. 63–69, 1998.

[8] Michele Geronazzo and Stefania Serafin, *Sonic Interactions in Virtual Environments*, Springer Nature, 2023.

[9] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell, "Active learning with gaussian processes for object categorization," in *2007 IEEE 11th international conference on computer vision*. IEEE, 2007, pp. 1–8.

[10] Benjamin Letham, Phillip Guan, Chase Tymms, Eytan Bakshy, and Michael Shvartsman, "Look-ahead acquisition functions for bernoulli level set estimation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 8493–8513.

[11] Manuel Haussmann, Fred A Hamprecht, and Melih Kandemir, "Deep active learning with adaptive acquisition," *arXiv preprint arXiv:1906.11471*, 2019.

[12] Alan B Cobo-Lewis, "An adaptive psychophysical method for subject classification," *Perception & Psychophysics*, vol. 59, no. 7, pp. 989–1003, 1997.

[13] Claude E Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[14] Carl Edward Rasmussen, Christopher KI Williams, et al., *Gaussian processes for machine learning*, vol. 1, Springer, 2006.

[15] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel, "Bayesian active learning for classification and preference learning," *arXiv preprint arXiv:1112.5745*, 2011.

[16] Mijung Park, Greg Horwitz, and Jonathan Pillow, "Active learning of neural response functions with gaussian processes," *Advances in neural information processing systems*, vol. 24, 2011.

[17] Bert de Vries, "A gaussian process mixture prior for hearing loss modeling," in *Benelearn 2017: Proceedings of the Twenty-Sixth Benelux Conference on Machine Learning, Technische Universiteit Eindhoven, 9-10 June 2017*, 2017, p. 74.

[18] Dennis L Barbour, Rebecca T Howard, Xinyu D Song, Nikki Metzger, Kiron A Sukesan, James C DiLorenzo, Braham RD Snyder, Jeff Y Chen, Eleanor A Degen, Jenna M Buchbinder, et al., "Online machine learning audiometry," *Ear and hearing*, vol. 40, no. 4, pp. 918, 2019.

[19] Jens Brehm Bagger Nielsen, Jakob Nielsen, and Jan Larsen, "Perception-based personalization of hearing aids using gaussian processes and active learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 162–173, 2014.

[20] Josef Schlittenlacher, Richard E Turner, and Brian CJ Moore, "Audiogram estimation using bayesian active learning," *The Journal of the Acoustical Society of America*, vol. 144, no. 1, pp. 421–430, 2018.

[21] Xinyu D Song, Brittany M Wallace, Jacob R Gardner, Noah M Ledbetter, Kilian Q Weinberger, and Dennis L Barbour, "Fast, continuous audiogram estimation using machine learning," *Ear and hearing*, vol. 36, no. 6, pp. e326, 2015.

[22] Josef Schlittenlacher and Brian CJ Moore, "Fast estimation of equal-loudness contours using bayesian active learning and direct scaling," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 358–360, 2020.

[23] Xinyu D Song, Roman Garnett, and Dennis L Barbour, "Psychometric function estimation by probabilistic classification," *The Journal of the Acoustical Society of America*, vol. 141, no. 4, pp. 2513–2525, 2017.

[24] Keith A May and Joshua A Solomon, "Four theorems on the psychometric function," *PLoS One*, vol. 8, no. 10, pp. e74815, 2013.

[25] James Hensman, Alex Matthews, and Zoubin Ghahramani, "Scalable variational gaussian process classification," *arXiv preprint arXiv:1411.2005*, 2014.

[26] David M Blei, Alp Kucukelbir, and Jon D McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[27] Thomas M Cover, *Elements of information theory*, John Wiley & Sons, 1999.

[28] Mathias Dietz, Stephan D Ewert, and Volker Hohmann, "Lateralization based on interaural differences in the second-order amplitude modulator," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 398–408, 2012.

[29] Felix A Wichmann and N Jeremy Hill, "The psychometric function: I. fitting, sampling, and goodness of fit," *Perception & psychophysics*, vol. 63, no. 8, pp. 1293–1313, 2001.

[30] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi, "Torchaudio: Building blocks for audio and speech processing," *arXiv preprint arXiv:2110.15018*, 2021.

[31] Miguel A Garcıa-Pérez, "Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties," *Vision research*, vol. 38, no. 12, pp. 1861–1881, 1998.

[32] William A Yost, Robert Turner, and Byron Bergert, "Comparison among four psychophysical procedures used in lateralization," *Perception & Psychophysics*, vol. 15, no. 3, pp. 483–487, 1974.

[33] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[34] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson, "Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration," *Advances in neural information processing systems*, vol. 31, 2018.

[35] Beverly A Wright and Matthew B Fitzgerald, "Different patterns of human discrimination learning for two interaural cues to sound-source location," *Proceedings of the National Academy of Sciences*, vol. 98, no. 21, pp. 12307–12312, 2001.

[36] Kourosh Saberi, "Some considerations on the use of adaptive methods for estimating interaural-delay thresholds," *The Journal of the Acoustical Society of America*, vol. 98, no. 3, pp. 1803–1806, 1995.

[37] Jennifer E Mossop and John F Culling, "Lateralization of large interaural delays," *The Journal of the Acoustical Society of America*, vol. 104, no. 3, pp. 1574–1579, 1998.

[38] Christopher KI Williams and David Barber, "Bayesian classification with gaussian processes," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998.

[39] Thomas P Minka, "Expectation propagation for approximate bayesian inference," *arXiv preprint arXiv:1301.2294*, 2013.

[40] Julien Bect, David Ginsbourger, Ling Li, Victor Picheny, and Emmanuel Vazquez, "Sequential design of computer experiments for the estimation of a probability of failure," *Statistics and Computing*, vol. 22, pp. 773–793, 2012.

[41] Mohammad Khan, *Variational learning for latent Gaussian model of discrete data*, Ph.D. thesis, University of British Columbia, 2012.

[42] Andrew B Watson and Denis G Pelli, "Quest: A bayesian adaptive psychometric method," *Perception & psychophysics*, vol. 33, no. 2, pp. 113–120, 1983.

[43] Josef Schlittenlacher, Richard E Turner, and Brian CJ Moore, "The implementation of efficient hearing tests using machine learning," in *Proceedings of the International Symposium on Auditory and Audiological Research*, 2019, vol. 7, pp. 1–12.

[44] Felix A Wichmann and N Jeremy Hill, "The psychometric function: Ii. bootstrap-based confidence intervals and sampling," *Perception & psychophysics*, vol. 63, pp. 1314–1329, 2001.

[45] Harry CCH Levitt, "Transformed up-down methods in psychoacoustics," *The Journal of the Acoustical society of America*, vol. 49, no. 2B, pp. 467–477, 1971.

[46] Brent Bryan, Robert C Nichol, Christopher R Genovese, Jeff Schneider, Christopher J Miller, and Larry Wasserman, "Active learning for identifying function threshold boundaries," *Advances in neural information processing systems*, vol. 18, 2005.

[47] Lucy Owen, Jonathan Browder, Benjamin Letham, Gideon Stocek, Chase Tymms, and Michael Shvartsman, "Adaptive nonparametric psychophysics," *arXiv preprint arXiv:2104.09549*, 2021.