Psychophysical Detection Testing with Bayesian Active Learning

Jacob R. Gardner
gardner.jake@wustl.edu
Washington University in St. Louis
St. Louis, MO 63130

Xinyu Song
xinyu.song@wustl.edu
Washington University in St. Louis
St. Louis, MO 63130

Kilian Q. Weinberger
kilian@wustl.edu
Washington University in St. Louis
St. Louis, MO 63130

Dennis Barbour
dbarbour@wustl.edu
Washington University in St. Louis
St. Louis, MO 63130

John P. Cunningham
jpc2181@columbia.edu
Columbia University
New York, NY 10027

Abstract

Psychophysical detection tests are ubiquitous in the study of human sensation and the diagnosis and treatment of virtually all sensory impairments. In many of these settings, the goal is to recover, from a series of binary observations from a human subject, the latent function that describes the discriminability of a sensory stimulus over some relevant domain. The auditory detection test, for example, seeks to understand a subject's likelihood of hearing sounds as a function of frequency and amplitude. Conventional methods for performing these tests involve testing stimuli on a pre-determined grid. This approach not only samples at very uninformative locations, but also fails to learn critical features of a subject's latent discriminability function. Here we advance active learning with Gaussian processes to the setting of psychophysical testing. We develop a model that incorporates strong prior knowledge about the class of stimuli, we derive a sensible method for choosing sample points, and we demonstrate how to evaluate this model efficiently. Finally, we develop a novel likelihood that enables testing of multiple stimuli simultaneously. We evaluate our method in both simulated and real auditory detection tests, demonstrating the merit of our approach.

1 INTRODUCTION

Psychophysical tests are a fundamental tool for investigating human perception: does a particular stimulus produce sensation for a particular person? The most common form of psychophysical tests – *detection tests* – present n sensory stimuli to a subject, and ask for n binary reports as to whether each stimulus was detected or not. Detection tests exist for vision (Schiefer et al., 2005), pain (Carter and Shieh, 2009), and many other settings. Perhaps the most common example is audiometry (Carhart and Jerger, 1959; Don et al., 1978; Hughson and Westlake, 1944): a subject is presented with a sequence of n tones $\mathbf{x}_t \ \ \forall t = 1, ..., n$, where each tone $\mathbf{x}_t \in \mathbb{R}^2$ is a pure tone with a specific frequency (pitch) and intensity (volume). The subject reports an observation $y_t = 1$ if he/she heard the tone, and a $y_t = 0$ is concluded in the absence of a positive report. The purpose of the test is to infer, from this sequence of observations, the underlying *audiometric function* $g(\mathbf{x})$, a function that describes how likely the subject is to hear sounds over the domain of typical frequencies and intensities. There is substantial variability in each person's audiogram, particularly for those with partial, selective, or degenerative hearing loss (Gosztonyi Jr et al., 1971; Robinson, 1991; Schmuziger et al., 2004). Accurate estimates of audiograms are thus essential to understanding human audition, and to all medical studies and treatments of various forms of hearing loss.

A standard auditory detection test is carried out by playing an n-length sequence of pure tones on a pre-defined grid in frequency-intensity space. This approach, while simple, has several salient drawbacks that lead to an unnecessarily large n. First, a given tone is played multiple times, even if it is highly audible or highly inaudible. Second, information is not shared between previous outcomes. For example, human audition is monotonically increasing in intensity, but in the standard test, even if a particular frequency of sound is heard at a given intensity, tones with the same frequency but higher intensity will still be tested. Finally, owing to limitations on the size of sequence n, a standard detection test probes only six discrete frequencies (Madison et al., 2005). The coarseness of this grid can cause significant errors, as human hearing loss can span a range narrow enough to be entirely missed by these six frequencies (Jerger, 1960; Zhao et al., 2002; Zhao and Stephens, 1998). All of these issues, combined with the impracticality and burden to human subjects of a large n sequence, motivate an active learning approach.

Here we treat psychophysical detection tests as an active learning problem, extending and adapting recent work on active learning with Gaussian processes (GPs) (Garnett et al., 2013; Houlsby et al., 2011; Iwata et al., 2013). Our method addresses all the drawbacks of grid-sampling by performing Bayesian active learning of the audiometric function $g(\mathbf{x})$. Specifically, we place a GP prior on the latent audiogram $f(\mathbf{x})$, which we transform to a $[0, 1]$ valued quantity using a probit transformation (Kuss and Rasmussen, 2005), such that $g(\mathbf{x}) \approx \Phi(f(\mathbf{x}))$. We use this model to sequentially sample at each time step t the most informative next tones conditioned on the previous $t-1$ observations $y_1, ..., y_{t-1}$. This model significantly enhances the accuracy and efficiency of learning audiograms. Our work offers two main contributions:

1. We extend and adapt existing work on Bayesian optimization and active learning to the setting of psychophysical detection tests. We present a model that incorporates strong prior knowledge about the auditory stimulus space, and we present experimental results demonstrating the effectiveness of a Bayesian active learning approach.

2. We develop a novel ‘OR-channel’ likelihood that allows the query of *multiple tones simultaneously*. We analyze this likelihood in the active learning context, clarifying the non-obvious intuition for why and when such an approach can outperform single-tone queries.

We evaluate our algorithm on both simulated and real audiometric detection tests. Our active learning approach obtains finer grained estimates of the audiogram $g(\mathbf{x})$ with substantially fewer stimuli queries (lower n). We note that, in the remainder of this work (notably our experiments), we will continue to use the example and nomenclature of audiometry, though our algorithm is precisely equivalent for other psychophysical detection tests as well.

2 GAUSSIAN PROCESSES

Throughout this paper we will make extensive use of Gaussian processes (GPs). A GP is formally a prior over functions, $f \sim \mathcal{GP}(\mu_0(\cdot), k(\cdot, \cdot))$, parameterized by a mean function $\mu_0(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ and covariance function $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - \mu_0(\mathbf{x}))(f(\mathbf{x}') - \mu_0(\mathbf{x}'))]$.

For any set of n observations $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, the GP implies that their function values $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\top$ are jointly Gaussian distributed, $\mathbf{f} \sim \mathcal{N}(\mu(\mathbf{X}), \mathbf{K})$, where $\mathbf{K}$ defines the covariance $\mathbf{K}_{ij} = \mathrm{Cov}[f_i, f_j] = k(\mathbf{x}_i, \mathbf{x}_j)$. If we add a test point $\mathbf{x}^*$ with unknown function value f^* this distribution extends naturally by one dimension to

$$\begin{bmatrix} \mathbf{f} \\ f^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(\mathbf{X}) \\ \mu(\mathbf{x}^*) \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{k}^* \\ \mathbf{k}^{*\top} & k(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix} \right).$$

We can utilize standard Gaussian conditioning rules (Rasmussen and Williams, 2006) to derive the posterior distribution, $p(f^*|\mathbf{X}, \mathbf{f}, \mathbf{x}^*)$, which is Gaussian with mean and variance

$$\begin{aligned} \mu^*(\mathbf{x}^*) &= \mu_0(\mathbf{x}^*) + \mathbf{k}^{*\top}\mathbf{K}^{-1}(\mathbf{f} - \mu_0(\mathbf{x}^*)) & (1) \\ \sigma^{*2}(\mathbf{x}^*) &= k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^{*\top}\mathbf{K}^{-1}\mathbf{k}^*. & (2) \end{aligned}$$

Here $\mathbf{k}^* = [k(\mathbf{x}^*, \mathbf{x}_1), ..., k(\mathbf{x}^*, \mathbf{x}_n)]^\top$ denotes the kernel vector between the test input $\mathbf{x}^*$ and each training input.

In practice, we often do not observe f_i directly, but rather some dependent random variable y_i. A popular example is to assume additive Gaussian noise, $y_i = f_i + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$. In this setting, the distribution for f^* remains Gaussian, with a mean and variance similar to eqs. (1) and (2) (where $\mathbf{K}$ is replaced with $\mathbf{K} + \sigma_n^2\mathbf{I}$).

However, with most observation models, the posterior distribution of f^* conditioned on $\mathbf{y}$ is not Gaussian, and exact inference becomes impossible. Approximate inference may be performed using a Gaussian approximation to the likelihood (Kuss and Rasmussen, 2005; Minka, 2001). In particular, by using a Gaussian approximation to the likelihood, we recover the Gaussianity of the posterior. For a full treatment of Gaussian processes, see (Rasmussen and Williams, 2006).

Note that in many cases, our goal is to make predictions, for which we use the posterior predictive distribution–a distribution over y^*:

$$p(y^*|\mathbf{X}, \mathbf{y}, \mathbf{x}^*) = \int_{f^*} p(y^*|f^*)p(f^*|\mathbf{X}, \mathbf{y}, \mathbf{x}^*)df^*, \quad (3)$$

This distribution is typically not computable analytically. However, if the posterior distribution for f^* is Gaussian (e.g., because a Gaussian likelihood or Gaussian approximate likelihood was used), this integral can often be computed efficiently.

2.1 BAYESIAN ACTIVE LEARNING

The goal of Bayesian active learning is to sequentially choose samples so as to accurately model an unknown function $g(\cdot)$ with as few samples as possible. In the audiometric setting, $g(\mathbf{x})$ is the probability that the patient hears the tone $\mathbf{x}$. If we query whether the patient can hear a set of tones $\mathbf{X}$, we would like for our predictive posterior belief $p(y^*|\mathbf{X}, \mathbf{y}, \mathbf{x}^*)$ to match $g(\mathbf{x}^*)$ as *well* as possible and as *confidently* as possible. Suppose that at iteration $t < n$ the points $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_t]$ and corresponding labels $\mathbf{y}$ are known. Houlsby et al. (2011) propose to use mutual information,

$$I(\mathbf{f}, y_t|\mathbf{x}_t) = H\left[\mathbf{f}|\mathbf{X}, \mathbf{y}\right] - \mathbb{E}\left[H[\mathbf{f}|\mathbf{X}, \mathbf{y}, y_t]\right]_{p(y_t|\mathbf{X}, \mathbf{y}, \mathbf{x}_t)} \quad (4)$$

where $H[A]$ denotes the differential entropy of a random variable A, to identify a new point $\mathbf{x}_t$, with future label y_t, to be queried in iteration t—*i.e.* $\mathbf{x}_t$ is chosen to be

$$\mathbf{x}_t = \arg\max_{\mathbf{x}} I(\mathbf{f}, y|\mathbf{x}) \quad (5)$$

3 METHOD

In this section, we discuss our model and approach to psychophysical detection testing using Gaussian processes. As a running example, we will use audiometry. In an audiometric detection test, a patient is presented with tones of varying frequency and intensity. The patient is asked to respond (*e.g.,* by pressing a button) if he/she hears the sound. In the absence of a timely reaction the tone is assumed to be inaudible to the patient. The delay between tones is sufficiently randomized to prevent patients from responding to predictable patterns (Gosztonyi Jr et al., 1971).

At time step t, we choose a tone $\mathbf{x}_t = (\omega, i)$ with frequency ω and intensity i to present to the subject. In return, we receive a response $y_t \in \{0, 1\}$, where $y_t = 1$ indicates that the patient heard the sound and $y_t = 0$ indicates that he/she did not. There is inherent observation noise in patient responses. When patients become uncertain when presented with sounds very close to their threshold (*i.e.,* the sounds become faint and hard to hear). Patients do not have perfect detection boundaries, and only hear tones near their hearing threshold with some probability. This uncertainty is observed in reality for a number of reasons. First, patient attention may waver, or they may be unable to distinguish between tones near their hearing threshold and slight background noise. Alternatively, this uncertainty may derive from physical sources. For example, if a tone is faint enough, a patient may be able to hear that tone between–but not during–heart beats. Our goal is therefore to predict the probability that a patient is able hear a given sound.

3.1 PRIOR

In the case of audiometric testing, we have valuable prior knowledge about a patient's audiometric function that we can encode in our GP model. In particular, the probability that a patient hears a sound (ω, i) is *monotonically increasing* in the intensity i. In other words, if a tone is audible to a patient, then an even louder tone is more likely to be audible. Furthermore, audition is a *smooth function* with respect to the frequency ω. Human nerves that detect similar frequencies are co-located in the cochlea and, as a result, a partial loss of hearing in one frequency is likely to cause a loss of hearing in nearby frequencies. A GP prior can encode both properties naturally through its covariance function. A combination of a linear kernel in intensity and a squared exponential kernel in frequency ensures the monotonicity and smoothness properties:

$$k\left((\omega, i), (\omega', i')\right) = ii' + \exp\left\{-\frac{1}{\ell}\|\omega - \omega'\|_2^2\right\}. \quad (6)$$

Here, ℓ regulates the smoothness (characteristic length-scale) w.r.t. frequency. Note that a GP prior is technically incapable of supporting only monotonically increasing functions. However, we only need that the posterior probability of detection, 3, be monotonic, which is generally true after a few tones are sampled (for example, see figure 3).

For the mean function μ_0, we note that intensity is typically measured in dB HL, which is an empirical unit of measurement normalized based on population data so that at each frequency the typical human hearing threshold is around 0 dB HL. As a result we choose a constant mean function.

3.2 OBSERVATION MODEL

This mean function, $\mu_0(\cdot)$, and covariance function, $k(\cdot, \cdot)$, define a prior over real-valued latent functions $f \sim \mathcal{GP}(\mu_0(\cdot), k(\cdot, \cdot))$. Our goal is to predict the *probability* (*i.e.* within $[0, 1]$) that a patient hears a tone with a specified frequency and intensity. We can never observe these probabilities directly. For any tone, we can instead only observe the outcome of a Bernoulli trial with the true probability. This setting is akin to Gaussian Process classification (Kuss and Rasmussen, 2005) and similarly we use a Bernoulli likelihood, where $\Pr(y = 1|f) = \Phi(f)$ and $\Phi(\cdot)$ denotes the standard normal cumulative density function (CDF).

The linear component of the kernel in (6) results in a function that, after being warped by $\Phi(\cdot)$, is sigmoidal in the intensity dimension: after the slope is fixed (by conditioning on the first few points), the posterior belief about $\Phi(f)$ will tend to 0 as the intensity decreases and 1 as the intensity increases. This reflects our prior knowledge that tones of extremely low intensity are unlikely to be heard, whereas tones of high intensity are more likely to be audible.

Predictions. Once we have collected data, we can use the predictive distribution $p(y^*|\mathbf{X}, \mathbf{y}\, \mathbf{x}^*)$ to summarize our belief about whether the patient will hear a test tone $\mathbf{x}^*$. As our likelihood is non-Gaussian, the posterior $p(f^*|\mathbf{X}, \mathbf{y}, \mathbf{x}^*)$ has no closed form solution. However, an approximate Gaussian posterior over f^* can be obtained with the standard Laplace approximation to the likelihood (Kuss and Rasmussen, 2005; Rasmussen and Williams, 2006).

3.3 MULTIPLE TONES

An interesting property of audiometry (that may also be common to other psychophysical domains, *e.g.* visual or touch sensory tests), is that multiple tone stimuli can be presented to a patient simultaneously by overlaying tones. In this setting however, we can still only query whether the patient heard the overlaid tones. A negative response to a multi-tone sample indicates that the patient did not hear any of the overlaid tones; a positive response indicates that the patient heard *at least one* of them.

OR-Channel. Presenting a patient with k tones leads to a novel extension to the standard Bernoulli likelihood used in classification. We present the patient with k tones $\mathbf{x}_1, ..., \mathbf{x}_k$. The patient hearing the individual tone $\mathbf{x}_i$ is still the outcome of a Bernoulli trial with $\Pr(y_i|f_i) = \Phi(f_i)$, as the individual trials are independent *conditioned* on f. However, we cannot directly observe any individual y_i. Rather, we record them through an *OR-channel*, that is we observe $\bar{y}$, which is 1 if the patient hears *at least one* of the k tones presented, and is 0 otherwise. This leads to the *OR-channel likelihood*:

$$\begin{aligned}\Pr(\bar{y} = 1|\mathbf{f}_{1..k}) &= 1 - \prod_j (1 - \Phi(f_j)) \\ &= 1 - \prod_j \Phi(-f_j) \qquad (7)\end{aligned}$$

Note when $k = 1$, eq. (7) reduces to the standard Bernoulli likelihood for single tones, $\Pr(\bar{y} = 1|f_1) = \Phi(f_1)$.

3.4 QUERY SELECTION

In iteration t we present the subject with a query set of overlaid tones $\mathbf{q}_t = [\{\mathbf{x}_1, ..., \mathbf{x}_k\}]$ and query the response $\bar{y}_t$. To select $\mathbf{q}_t$ we pick the point set that maximizes the expected decrease in posterior entropy, analogous to eq. (4).

Single tone mutual information. We first consider the setting of picking a single tone, *i.e.* where $\mathbf{q}_t = [\{\mathbf{x}_t\}]$. Houlsby et al. (2011) derive an analytical approximation to the mutual information, eq. (5), when using a Bernoulli likelihood. These results directly apply when picking a single tone $\mathbf{x}_t$. When f_t is known, the entropy of the Bernoulli variable y_t is given by $\mathrm{h}(\Phi(f_t))$, where

$$\mathrm{h}(p) = -p \log p - (1-p) \log(1-p),$$

is the Bernoulli entropy function. We can rephrase the entropy in eq. (4) as

$$I(\mathbf{f}, y_t|\mathbf{q}_t) = H\left[y_t|\mathbf{X}, \mathbf{y}\right] - \mathbb{E}\left[H\left[y_t|\mathbf{f}\right]\right]_{p(\mathbf{f}|\mathbf{X},\mathbf{y})}, \qquad (8)$$

and rewrite both terms on the right hand side through h. If $\mathbf{f}$ is unknown and y_t is conditioned on $\mathbf{X}, \mathbf{y}$, the entropy can be expressed in terms of the expectation over the posterior for f_t:

$$H\left[y_t|\mathbf{X}, \mathbf{y}\right] = \mathrm{h}\left(\mathbb{E}\left[\Phi(f_t)\right]\right). \qquad (9)$$

If f_t is known we have $\Pr(y|\mathbf{f}) = \Phi(f_t)$, yielding

$$H\left[y_t|\mathbf{f}\right] = \mathrm{h}\left(\Phi(f_t)\right). \qquad (10)$$

Substituting eqs. (9), (10) into (8) leads us to the following expression for the mutual information between $\mathbf{f}$ and y_t in the single tone scenario:

$$I_1(\mathbf{f}, y_t|\mathbf{q}_t) = \mathrm{h}\left(\mathbb{E}\left[\Phi(f_t)\right]\right) - \mathbb{E}\left[\mathrm{h}\left(\Phi(f_t)\right)\right]. \qquad (11)$$

The computation of I_1 involves an intractable integral, which can be approximated through numerical integration. This approach is very fast in practice as the integral is only one dimensional and can be computed efficiently using quadrature.

Multiple tone mutual information The above results can be extended to compute the mutual information when sampling multiple tones $\mathbf{q}_t = [\{\mathbf{x}_1, ..., \mathbf{x}_k\}]$. In particular, the probability of observing $\bar{y}_t = 1$ changes from $\Phi(f_t)$ to the OR-channel probability, (7). Thus, when $f_1, ..., f_k$ are known, the entropy of the Bernoulli variable $\bar{y}_t$ is $\mathrm{h}\left(1 - \prod_{i=1}^k \Phi(-f_i)\right)$.

To simplify notation, let us define $\bar{p}_1 = \Pr(\bar{y} = 1|\mathbf{f}_{1..k})$ as defined in (7). Substituting $\bar{p}_1$ for $\Phi(f_t)$ in (11) gives the mutual information of paired tone sample $\mathbf{q}_t$ after observing the outcome $\bar{y}_t$:

$$\begin{aligned}I_k(\mathbf{f}, \bar{y}_t|\mathbf{q}_t) &= \mathrm{h}(\mathbb{E}\left[\bar{p}_1\right]) - \mathbb{E}\left[\mathrm{h}(\bar{p}_1)\right] \\ &= \mathrm{h}\Big(\mathbb{E}\Big[\textstyle\prod_j \Phi(-f_j)\Big]\Big) - \mathbb{E}\Big[\mathrm{h}\Big(\textstyle\prod_j \Phi(-f_j)\Big)\Big] \qquad (12)\end{aligned}$$

where the second equality holds by the linearity of expectation and because $\mathrm{h}(p)$ is a concave function that is symmetric about $p = 0.5$ (*i.e.* $\mathrm{h}(p) = 1 - \mathrm{h}(p)$). The last term leads again to an intractable integral. However, similar to the one tone scenario, I_k can also be evaluated efficiently using numerical integration, as k is relatively small.

Computational Considerations Finding a set of $k \leq K$ tones $\mathbf{q}_t^{(k)}$ to maximize $I_k(\mathbf{f}, \bar{y}_t|\mathbf{q}_t)$ from a candidate set $\mathcal{X}$ of size S requires $O\left(\binom{S}{k}\right)$ considerations. In order to ensure that patients do not have to wait for a lengthy duration between sounds are played, we construct a set of multiple tones to play greedily. We select the best single tone by exhaustively searching $\mathcal{X}$. Then, to select the best set of size k, we exhaustively add each $\hat{\mathbf{x}} \in \mathcal{X}$ to the best set of size $k-1$, $\mathbf{q}_t^{(k-1)}$ and compute the expected decrease in posterior entropy of $\mathbf{q}_t^{(k-1)} \cup \hat{\mathbf{x}}$. This greedy selection procedure reduces the computational complexity of considering tone sets of up to size k to $O(Sk)$, and in practice requires only a few seconds of computation time.

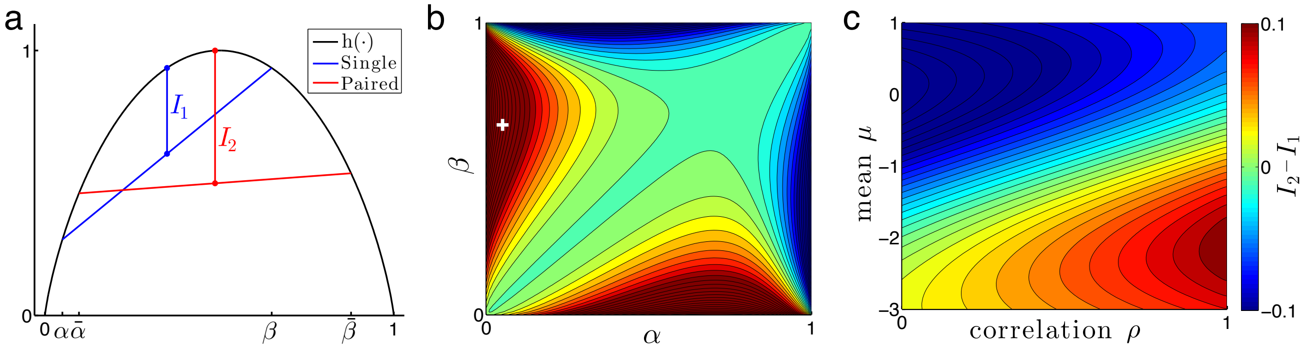


Figure 1: Difference in mutual information $I_2 - I_1$ between a paired query and a single query: **(a)** discrete distribution with two atoms $(\alpha, \beta) = 0.05, 0.65$, and corresponding $\bar{\alpha} = 1 - (1-\alpha)^2 \approx 0.1$, $\bar{\beta} \approx 0.88$). Here $I_2 - I_1 \approx 0.18$; **(b)** $I_2 - I_1$ as a function of α, β (white cross denotes the specific example of panel *a*); **(c)** the normally distributed latent input case. $I_2 - I_1$ is shown as a function of the mean μ and correlation ρ. Colorbar at right is for both panel *b* and *c*.

3.5 OR-CHANNEL ANALYSIS

We first investigate the OR-channel likelihood of eq. (7), as it is unclear if this elaboration can provide any benefit over a standard Bernoulli likelihood. Intuitively, the result of $\bar{y} = 0$ from an OR-channel is quite informative: all inputs into that channel must have been 0 (in the auditory example, no sounds were heard). On the other hand, the result of $\bar{y} = 1$ is much less informative than in the Bernoulli channel, as it means only that one or more of the inputs were 1 (some sound or sounds were heard), but there is no information about which. Here we analyze simple models that support the use of the OR-channel likelihood. We compare a *single* input, corresponding to the standard Bernoulli likelihood, to a *paired* input, corresponding to an OR-channel likelihood with two inputs. That is, with inputs $\{f_1, f_2\}$ and output $y \in \{0, 1\}$ as above, our quantities of interest are $I_1 := I(y, f_1)$ and $I_2 := I(y, \{f_1, f_2\})$, and we seek to understand if more information about the inputs can exist in the paired-input query, than in the single-input query.

3.5.1 OR-channel Inputs With Discrete Support

The simplest case involves perfectly correlated inputs $f_1 = f_2$, and further, a discrete distribution on f_1 with two atoms of equal mass. The implied probability $\phi(f_1)$ will then have the same discrete distribution, which we write as $p\left(\phi(f_1)\right) = \frac{1}{2}\delta(\phi(f_1) = \alpha) + \frac{1}{2}\delta(\phi(f_1) = \beta)$, for some atoms α and β. Then, the mutual information of the single query is:

$$\begin{aligned} I_1 &= H(y) - H(y|f_1) \\ &= \mathrm{h}\left(\mathbb{E}_f\left[\phi\left(f_1\right)\right]\right) - \mathbb{E}_f\left[\mathrm{h}\left(\phi\left(f_1\right)\right)\right] \\ &= \mathrm{h}\left(\frac{1}{2}(\alpha + \beta)\right) - \frac{1}{2}\left(\mathrm{h}(\alpha) + \mathrm{h}(\beta)\right), \end{aligned} \tag{13}$$

where $\mathbb{E}_f$ is the expectation under the distribution on f. The OR-channel likelihood for two terms is similarly $p(y = 1|\{f_1, f_2\}) = 1 - (1 - \phi(f_1))\,(1 - \phi(f_2)) = 1 - (1 - \phi(f_1))^2$. The mutual information of a paired-input query becomes

$$I_2 = \mathrm{h}\left(\frac{1}{2}\left(\bar{\alpha} + \bar{\beta}\right)\right) - \frac{1}{2}\left(\mathrm{h}\left(\bar{\alpha}\right) + \mathrm{h}\left(\bar{\beta}\right)\right), \tag{14}$$

where $\bar{\alpha} = 1 - (1 - \alpha)^2$ and $\bar{\beta} = 1 - (1 - \beta)^2$. I_2 and I_1 offer a convenient geometric interpretation by viewing mutual information as the Jensen's inequality gap of h (eqs. (13) and (14)). With this simple discrete distribution, α and β can be chosen such that $I_2 - I_1$ will be positive or negative. We show the critical case $I_2 > I_1$ in Figure 1a, where the blue line segment connects $(\alpha, \mathrm{h}(\alpha))$ to $(\beta, \mathrm{h}(\beta))$ with $(\alpha, \beta) = (0.05, 0.65)$, and the red line segment is then implied by those choices of α, β (that is, $(\bar{\alpha}, \bar{\beta}) \approx (0.10, 0.88)$ in the figure). Here the difference is $I_2 - I_1 = 0.18$ bits. The contours of $I_2 - I_1$ as a function of (α, β) is shown in Figure 1b.

3.5.2 OR-channel Inputs With Normal Densities

We next analyze the OR-channel likelihood with two latent factors $f_1 = f(x_1)$ and $f_2 = f(x_2)$, which are jointly Gaussian according to the GP model of Section 3: $[f_1, f_2] \sim \mathcal{N}(m, S)$. We calculate $I_2 - I_1$ numerically using eq. (11) (note that, compared to the previous example, only the expectation over f has changed). We simplify the parameter space with $m = \begin{bmatrix} \mu \\ \mu \end{bmatrix}$ and $S = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ (but note that the function $I_2 - I_1$ is not invariant to either of these simplifications). We plot the contours of $I_2 - I_1$ as a function of correlation ρ and mean μ in Figure 1c, which indeed has substantial regions of both positive and negative mass.

In summary, though intuitively non-obvious, the above analyses clarify that the OR-channel likelihood can, but need not, increase mutual information between the input distribution and the binary outcome y. This finding offers a

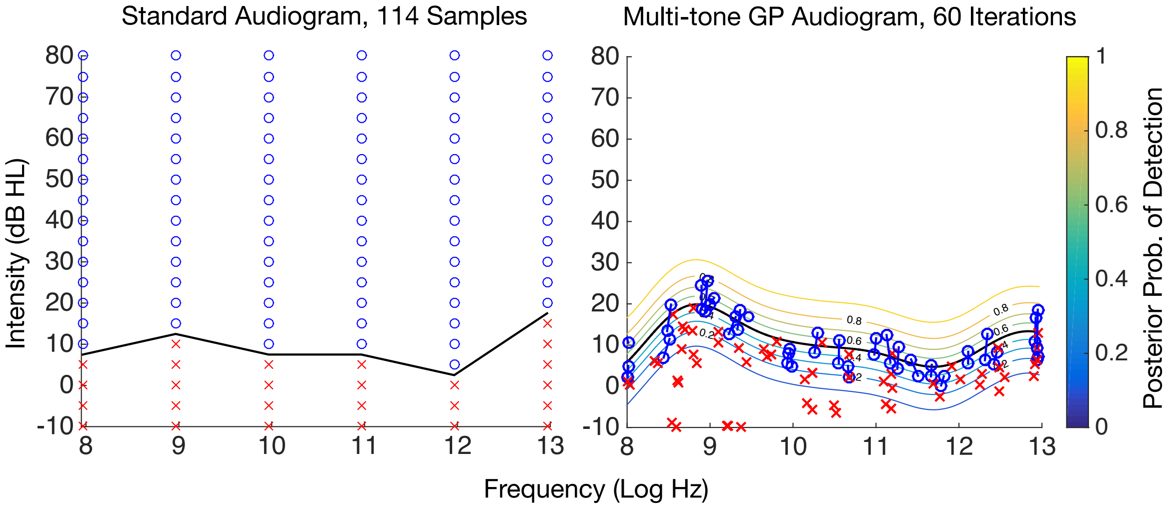


Figure 2: Standard grid search audiogram with tones played at every octave from 250 to 8000 Hz, and every 5 dB HL from -10 dB to 80 dB, compared to a multi-tone GP audiogram with 60 iterations (and therefore 119 "samples").

critical takeaway: the OR-channel can be used effectively, but only in the setting where a judicious choice of input distribution can be made. Indeed, this is exactly what our framework will achieve: it will choose pairs of input points (paired sounds) to learn more about the underlying audiogram than a single point alone. Thus, the OR-channel likelihood offers benefit beyond this scheme, which we already expect to outperform a naive approach to learning these latent functions. In this work we only consider paired inputs; a future question for study is how the information gain distribution changes with increasing numbers of inputs.

4 RELATED WORK

A number of papers have been recently published on Bayesian active learning. Many papers have considered Bayesian active learning using mutual information in the regression setting (Guestrin et al., 2005; Krause and Guestrin, 2007; Srinivas et al., 2009). However, the computation of mutual information is significantly less tractable in the classification setting. To our knowledge, Houlsby et al. (2011) is the first paper to leverage the rewriting of mutual information in (12), allowing for tractable computation of mutual information with the Bernoulli observation model. This paper is most similar to ours, as the Bernoulli observation model is identical to our single tone audiometric algorithm. A number of other, orthogonal applications and extensions of this method have since been published (Garnett et al., 2013; Iwata et al., 2013).

Alternative techniques for estimating audiograms have existed for many years. Sweep-based audiometry, such as Bekesy audiometry and Audioscan, are able to produce a more continuous estimate of the audiogram that can often detect notches, but with the disadvantage of a particularly time- and attention-demanding task (Jerger, 1960; Meyer-Bisch, 1996). Several Bayesian audiogram estimation techniques, such as parameter estimation by sequential testing (PEST) and maximum likelihood methods also exist, although most do not simultaneously estimate multiple frequencies (Green, 1993; Leek et al., 2000; Özdamar et al., 1990; Pentland, 1980; Taylor and Creelman, 1967). More recent advances in audiometric testing have focused on improving the accessibility of hearing screening by distribution over telephone, Internet, or mobile devices (Smits et al., 2004; Swanepoel et al., 2014; Vlaming et al., 2014; Watson et al., 2012; Williams-Sanchez et al., 2014).

5 RESULTS

In this section, we empirically evaluate our proposed algorithms for psychophysical detection. We focus on our application to audiometry, and seek to evaluate the merits of using Gaussian processes for audiometry in general, as well as to compare single-tone and multi-tone audiometry, focusing on the machine learning aspects of our algorithms.

We have since published a small clinical trial in a medical journal evaluating the novel GP audiometric techniques discussed here from a clinical point of view as well, and refer readers to Song et al. (2015) for additional results comparing GP audiometry and standard audiometry.

To begin, we compare the audiograms found by a standard grid audiometric test and by our multi-tone GP model. In both cases we run the same human subject in the same audiometric setting. The only differences are the tones presented and the method used to infer the audiometric function. All audiometric tests were run in accordance with an approved IRB. In the standard setting, tones from this

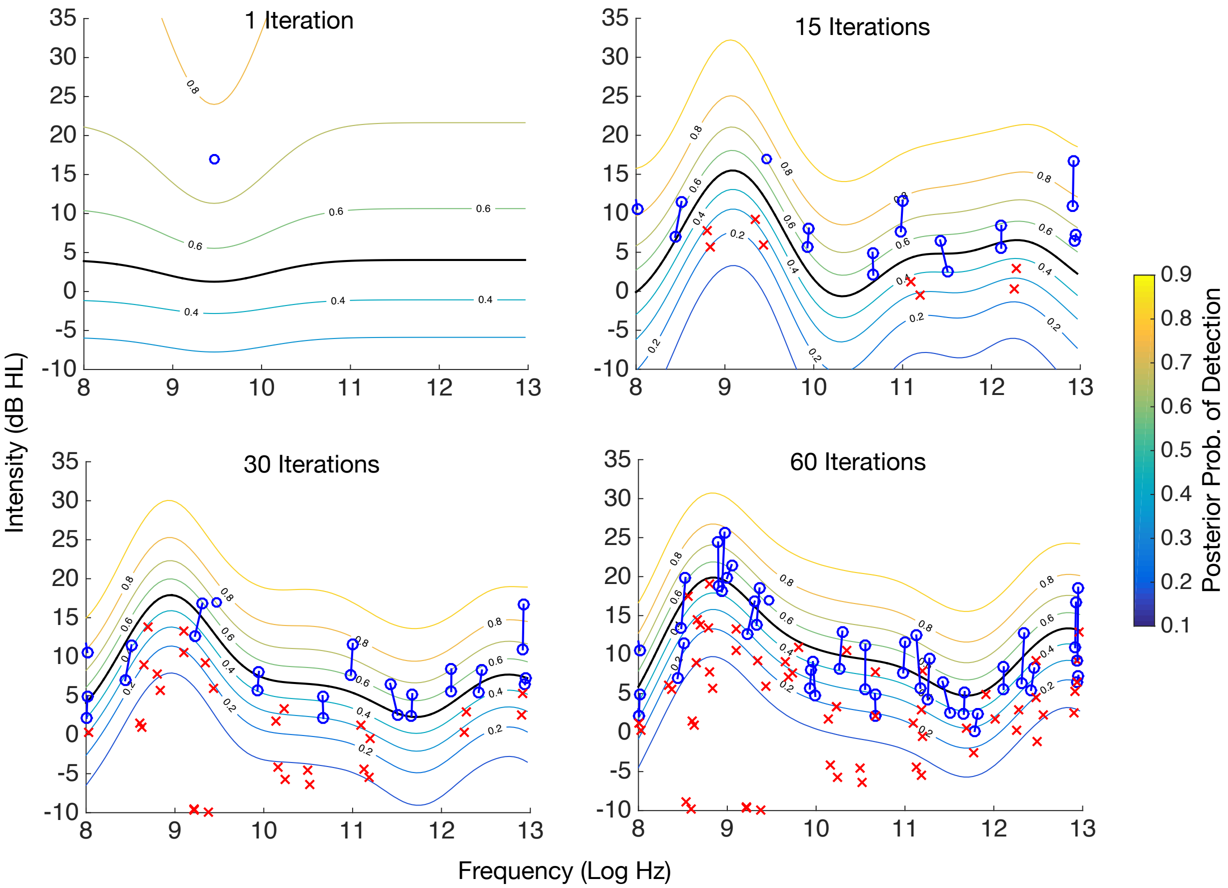


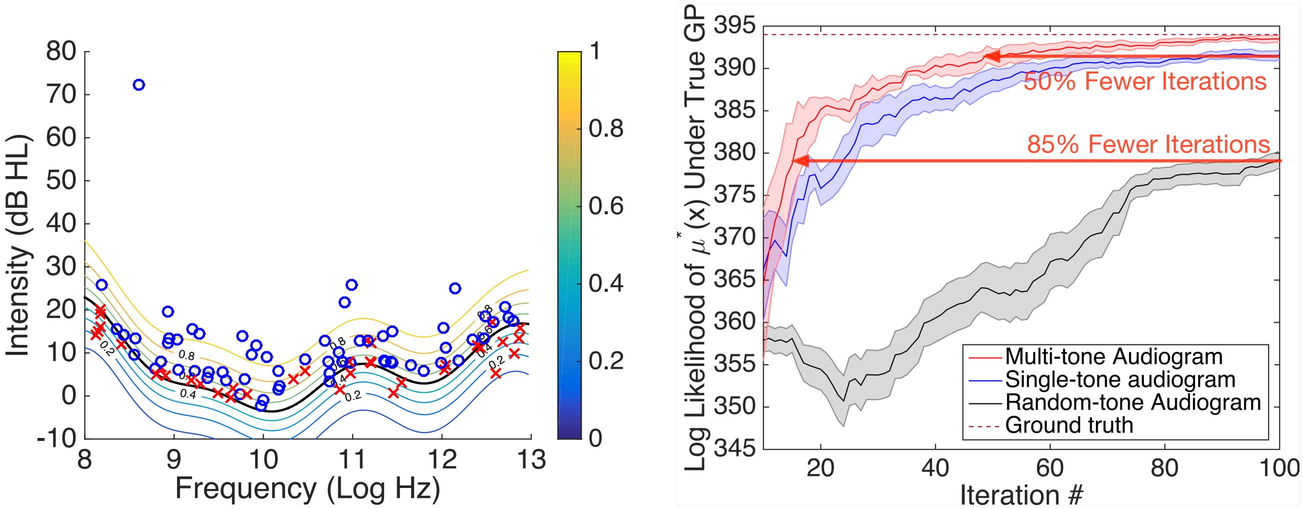
Figure 3: The posterior probability of detection within the frequency / intensity space during a GP audiometric test on a human subject. Panels show the learned GP after 1, 15, 30, and 60 iterations. Queries consist of a single or a paired tone (as selected by the model). Blue circles indicate a positive outcome (sound was heard), red crosses indicate a negative outcome. Paired tones with positive outcome (at least one of the two tones was heard) are connected by a blue line. Almost all queries are close to the final audible threshold (0.5 posterior detection probability), which is well approximated even after only 15 iterations.

grid are presented in a pre-determined order, typically ascending in frequency and decreasing in intensity. In the GP model, pairs of tones were actively selected given all previous pairs of tones and the responses to those tones. A random delay of up to 3 seconds was inserted between tone presentations to prevent subjects from memorizing a pattern in the test. Figure 2 shows the resulting data and inferred audiograms plotted in frequency-intensity space (left panel: standard audiometric test; right panel: GP method).

For both the standard and GP experiments, tones that were detected by the patient are plotted as blue circles, and tones that were not detected are plotted as red crosses. For the paired-tone GP test (right panel), paired samples that were detected are plotted as blue circles connected by a blue line (recall that, due to the OR-channel likelihood, we do not know which tone was heard). Paired tones that were not detected are again plotted as individual red crosses, as these data are functionally equivalent to two single-tone samples that were not detected (again due to the OR-channel observation model).

In the standard audiometric test, the inferred audiogram is simply an "audible threshold" that is the piecewise linear function connecting the detection threshold at each frequency. This threshold is depicted as a black line in the left panel of Figure 2. In the GP case, we infer a full posterior distribution on the detection threshold. We plot contours of the posterior detection probability in the right panel of Figure 2, with a solid black line at 50% posterior detection probability.

This confirmatory comparison offers several key points of interpretation. First, the tests agree with each other: the 50% posterior detection probability in the GP case is within 5dB of the standard audiogram, giving confidence to the general sensibility of this model. Second, perhaps most importantly to the active learning goal, the GP active learning



(a) A GP trained on 100 single tones. Blue circles denote tones detected by the subject, and red crosses denote tones that were not detected. The posterior probabilities are shown as color contours.

(b) Log likelihood of random presentation of tones (no active learning, shown in gray), active learning presentation of single tones (shown in blue), and active learning with paired tones (shown in red), under the ground truth audiometric function from Figure 4a. Log likelihood is plotted as a function of iterations in each audiometric testing strategy. Shaded areas denote standard error.

Figure 4: Comparison of multi-tone and single-tone GP audiometrics

model presents approximately half as many iterations (60 actively learned paired tones compared to 114 single tones preselected from a grid). Thus the GP model is able to explore substantially more of the frequency space than the standard grid test, and it does so in many fewer overall iterations, reducing the burden of these tests. Third, note that the GP model does not explore uninformative regions of tone space: above a certain intensity (at which the model is confident that tones are certainly heard), there are no tones queried. This observation differs sharply from the standard test, which squanders numerous samples at intensities well above this subject's audible threshold, where little to no information is available. Fourth, by design our GP model offers a full posterior distribution over tone space, and thus produces a richer and more descriptive audiogram than the piecewise linear audible threshold function in the standard test. Finally, it is worth noting that, though the paired tones in the right panel of Figure 2 appear to be sampled at very similar frequencies in log-space, the differences were often nontrivial, up to four or five half steps in an octave.

Next, Figure 3 investigates the convergence of our GP model after $1, 15, 30, 60$ iterations of our paired-tone GP audiometric algorithm. The posterior after a single iteration (upper left panel) reflects primarily the prior mean and the covariance of the model, which incorporates our knowledge about the general shape of human audiograms. As the active learning procedure continues (other panels), the GP posterior quickly converges to the audiogram of this particular subject. After only 30 iterations, the GP model has already captured the audiogram shape, and subsequent changes are very minor.

To investigate the performance of our GP active learning method in greater detail, we construct a synthetic data set with known ground truth (a known audiometric function). We begin by training a GP on 100 single tones and the detection of those tones reported by a second human subject. The tones sampled and the inferred audiogram are presented in Figure 4a. We use this posterior GP as the true audiogram of a simulated subject.

This ground truth audiometric function allows for the critical assessment of performance shown in Figure 4b. We compare three strategies of data presentation: random presentation of tones (no active learning, shown in gray), active learning presentation of single tones (shown in blue), and active learning with paired tones (shown in red). For each strategy, at each iteration (tone presentation), we infer the GP posterior mean, which is the MAP estimate of the audiometric function, given each stream of data. We evaluate the log likelihood of each strategy's GP posterior mean under the ground truth GP from Figure 4a. This step offers a quantitative assessment of how closely each strategy has approximated the true audiometric function. The maroon dashed line depicts the log likelihood of the ground truth GP itself, which is thus the maximum achievable performance of any strategy. All three strategies (random, single tone active learning, paired tone active learning) should, with enough iterations, converge to ground truth. Thus, the essential question of this work, and indeed of any active learning method, is how much more quickly a particular strategy approaches the ground truth than competing strategies.

We ran the single and paired tone active learning methods ten times each, and standard errors are plotted as shaded regions. Because of the very high standard error of the random tone audiogram, these results were averaged over 100 runs.

Figure 4b has a few key findings. Both the single and paired tone active learning strategies significantly outperform random sampling. Thus our strong prior rapidly learns that large portions of the tone space are either very likely or very unlikely to be heard, and is able to quickly learn to sample in regions of high information. After 80-90 iterations the paired tone algorithm matches the ground truth model very closely. This result is in significant contrast to randomly choosing tones, which not only has very large standard error, but also rarely converges to a good model. Finally, we observe that the paired tone active learning strategy significantly outperforms the single tone strategy. In fact, the paired tone strategy requires only half as many iterations to achieve the same level of likelihood. Compared to random sampling, paired tone active learning reduces the number of iterations by 85%.

6 DISCUSSION

In this paper, we explored the problem of adapting Bayesian active learning to psychophysical testing, and improving upon standard techniques used in audiometric testing. In the process of our investigation, we developed a novel OR-channel likelihood that allows us to present multiple tones to a subject simultaneously, leading to an audiometric testing strategy that not only yields good audiogram estimation using significantly fewer samples, but also leads to much better coverage of the frequency dimension. We demonstrate a non-obvious result, that multiple tones played through an OR-channel can, but do not have to, yield more information than a single tone. As future work we will continue to investigate the theoretical properties of this likelihood function and its use in active learning. We also hope that the drastic improvements of our method over the state-of-the-art will convince experts in medicine and psychology to adapt machine learned approaches for psychophysical testing.

7 ACKNOWLEDGEMENTS

KQW and JRG are supported by NIH grant U01 1U01NS073457-01 and NSF grants IIA-1355406, IIS-1149882, EFRI-1137211, CNS-1017701, CCF-1215302, and IIS-1343896. XS and DB are supported by NIH grant R01-DC009215. JPC is supported by a Sloan Research Fellowship.

References

Raymond Carhart and James Jerger. Preferred method for clinical determination of pure-tone thresholds. *Journal of Speech & Hearing Disorders*, 1959.

Matt Carter and Jennifer C Shieh. *Guide to research techniques in neuroscience*. Academic Press, 2009.

Manuel Don, Jos J Eggermont, and Derald E Brackmann. Reconstruction of the audiogram using brain stem responses and high-pass noise masking. *The Annals of otology, rhinology & laryngology. Supplement*, (3 Pt 2 Suppl 57):1–20, 1978.

Roman Garnett, Michael A Osborne, and Philipp Hennig. Active learning of linear embeddings for gaussian processes. *arXiv preprint arXiv:1310.6740*, 2013.

Rudolph E Gosztonyi Jr, Lawrence A Vassallo, and Joseph Sataloff. Audiometric reliability in industry. *Archives of Environmental Health: An International Journal*, 22(1): 113–118, 1971.

David M Green. A maximum-likelihood method for estimating thresholds in a yes–no task. *The Journal of the Acoustical Society of America*, 93(4):2096–2105, 1993.

Carlos Guestrin, Andreas Krause, and Ajit Paul Singh. Near-optimal sensor placements in gaussian processes. In *ICML*, 2005.

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

WAITER Hughson and Harold Westlake. Manual for program outline for rehabilitation of aural casualties both military and civilian. *Trans Am Acad Ophthalmol Otolaryngol*, 48(Suppl):1–15, 1944.

Tomoharu Iwata, Neil Houlsby, and Zoubin Ghahramani. Active learning for interactive visualization. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 342–350, 2013.

James Jerger. Bekesy audiometry in analysis of auditory disorders. *Journal of Speech, Language, and Hearing Research*, 3(3):275–287, 1960.

Andreas Krause and Carlos Guestrin. Nonmyopic active learning of gaussian processes: an exploration-exploitation approach. In *ICML 24*, 2007.

Malte Kuss and Carl Edward Rasmussen. Assessing approximate inference for binary gaussian process classification. *The Journal of Machine Learning Research*, 6: 1679–1704, 2005.

Marjorie R Leek, Judy R Dubno, Ning-ji He, and Jayne B Ahlstrom. Experience with a yes–no single-interval maximum-likelihood procedure. *The Journal of the Acoustical Society of America*, 107(5):2674–2684, 2000.

Ted Madison et al. Guidelines for manual pure-tone threshold audiometry. 2005.

Christian Meyer-Bisch. Audioscan: a high-definition audiometry technique based on constant-level frequency sweeps-a new method with new hearing indicators. *International Journal of Audiology*, 35(2):63–72, 1996.

Thomas P Minka. Expectation propagation for approximate bayesian inference. In *UAI*, 2001.

Özcan Özdamar, Rebecca E Eilers, Edward Miskiel, and Judith Widen. Classification of audiograms by sequential testing using a dynamic bayesian procedure. *The Journal of the Acoustical Society of America*, 88(5): 2171–2179, 1990.

Alex Pentland. Maximum likelihood estimation: The best pest. *Attention, Perception, & Psychophysics*, 28(4): 377–379, 1980.

C.E. Rasmussen and C.K.I. Williams. Gaussian processes for machine learning. MIT Press, 2006.

DW Robinson. Long-term repeatability of the pure-tone hearing threshold and its relation to noise exposure. *British journal of audiology*, 25(4):219–235, 1991.

U Schiefer, J Pätzold, and F Dannheim. Konventionelle perimetrie. *Der Ophthalmologe*, 102(6):627–646, 2005.

Nicolas Schmuziger, Rudolf Probst, and Jacek Smurzynski. Test-retest reliability of pure-tone thresholds from 0.5 to 16 khz using sennheiser hda 200 and etymotic research er-2 earphones. *Ear and hearing*, 25(2):127–132, 2004.

Cas Smits, Theo S Kapteyn, and Tammo Houtgast. Development and validation of an automatic speech-in-noise screening test by telephone. *International journal of audiology*, 43(1):15–28, 2004.

X. D. Song, B. M. Wallace, J. R. Gardner, N. M. Ledbetter, K. Q. Weinberger, and D. L Barbour. Fast, continuous audiogram estimation using machine learning. *Ear and Hearing*, 2015.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

De Wet Swanepoel, Hermanus C Myburgh, David M Howe, Faheema Mahomed, and Robert H Eikelboom. Smartphone hearing screening with integrated quality control and data management. *International journal of audiology*, 53(12):841–849, 2014.

MiM Taylor and C Douglas Creelman. Pest: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America*, 41(4A):782–787, 1967.

Marcel SMG Vlaming, Robert C MacKinnon, Marije Jansen, and David R Moore. Automated screening for high-frequency hearing loss. *Ear and hearing*, 35(6): 667, 2014.

Charles S Watson, Gary R Kidd, James D Miller, Cas Smits, and Larry E Humes. Telephone screening tests for functionally impaired hearing: Current use in seven countries and development of a us version. *Journal of the American Academy of Audiology*, 23(10):757–767, 2012.

Victoria Williams-Sanchez, Rachel A McArdle, Richard H Wilson, Gary R Kidd, Charles S Watson, and Andrea L Bourne. Validation of a screening test of auditory function using the telephone. *Journal of the American Academy of Audiology*, 25(10):937–951, 2014.

F Zhao, D Stephens, and C Meyer-Bisch. The audioscan: a high frequency resolution audiometric technique and its clinical applications. *Clinical Otolaryngology & Allied Sciences*, 27(1):4–10, 2002.

Fei Zhao and Dafydd Stephens. Analyses of notches in audioscan and dpoaes in subjects with normal hearing. *International Journal of Audiology*, 37(6):335–343, 1998.