Robust and efficient Bayesian adaptive psychometric function estimation

Clement S. J. Doire, Mike Brookes, and Patrick A. Naylor

Citation: The Journal of the Acoustical Society of America **141**, 2501 (2017); doi: 10.1121/1.4979580
View online: http://dx.doi.org/10.1121/1.4979580
View Table of Contents: http://asa.scitation.org/toc/jas/141/4
Published by the Acoustical Society of America

Articles you may be interested in

Psychometric function estimation by probabilistic classification
The Journal of the Acoustical Society of America **141**, (2017); 10.1121/1.4979594

How many images are in an auditory scene?
The Journal of the Acoustical Society of America **141**, (2017); 10.1121/1.4981118

Effects of hearing-aid dynamic range compression on spatial perception in a reverberant environment
The Journal of the Acoustical Society of America **141**, (2017); 10.1121/1.4979783

Streaming and sound localization with a preceding distractor
The Journal of the Acoustical Society of America **141**, (2017); 10.1121/1.4979167

Robust source localization from wavefield separation including prior information
The Journal of the Acoustical Society of America **141**, (2017); 10.1121/1.4979258

Pupillometry shows the effort of auditory attention switching
The Journal of the Acoustical Society of America **141**, (2017); 10.1121/1.4979340

Robust and efficient Bayesian adaptive psychometric function estimation

Clement S. J. Doire, Mike Brookes,[a] and Patrick A. Naylor
Department of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, United Kingdom

(Received 1 July 2016; revised 16 March 2017; accepted 20 March 2017; published online 10 April 2017)

The efficient measurement of the threshold and slope of the psychometric function (PF) is an important objective in psychoacoustics. This paper proposes a procedure that combines a Bayesian estimate of the PF with either a look one-ahead or a look two-ahead method of selecting the next stimulus presentation. The procedure differs from previously proposed algorithms in two respects: (i) it does not require the range of possible PF parameters to be specified in advance and (ii) the sequence of probe signal-to-noise ratios optimizes the threshold and slope estimates at a performance level, ϕ, that can be chosen by the experimenter. Simulation results show that the proposed procedure is robust and that the estimates of both threshold and slope have a consistently low bias. Over a wide range of listener PF parameters, the root-mean-square errors after 50 trials were $\sim$1.2 dB in threshold and 0.14 in log-slope. It was found that the performance differences between the look one-ahead and look two-ahead methods were negligible and that an entropy-based criterion for selecting the next stimulus was preferred to a variance-based criterion. *© 2017 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).*
[http://dx.doi.org/10.1121/1.4979580]

[ICB] Pages: 2501–2512

I. INTRODUCTION

The performance of a listener on a psychoacoustic task is conveniently characterized by the psychometric function (PF), which gives the listener's response as a function of a stimulus parameter such as intensity or signal-to-noise ratio (SNR). For example, a widely used PF in the field of speech intelligibility expresses the fraction of words that are correctly identified in a noisy speech sample as a function of the SNR of the speech sample (Miller *et al.*, 1951). The primary goal of a speech intelligibility test is generally to measure the speech-reception threshold (SRT_ϕ), i.e., the SNR at which a particular fraction, ϕ, of words are correctly identified (Plomp and Mimpen, 1979; Wilson *et al.*, 2007). It is also, however, often important to measure the slope of the PF at the SRT (MacPherson and Akeroyd, 2014; Neuman *et al.*, 2010) and it is this that determines the perceptual benefit that can be obtained by improving the SNR.

This paper presents an adaptive procedure for measuring both the SRT and the slope of the PF. The procedure differs from previously proposed algorithms in two respects: (i) it does not require the range of possible PF parameters to be specified in advance and (ii) the sequence of probe SNRs is selected to optimize the estimates of the SRT and the slope of the PF at a performance level, ϕ, that can be chosen by the experimenter. The procedure makes efficient use of a limited number of measurements, has low bias, and is robust to listener variation. The primary application is in speech intelligibility testing but it is also applicable to other domains for which a suitable monotonic PF model exists.

A. PF models

A number of sigmoid-shaped functions have been used to provide a parametric model of the PF including the Weibull, Logistic, cumulative Gaussian, and Gumbel distribution (Wichmann and Hill, 2001). A common choice for speech intelligibility is the four-parameter transformed Logistic model (Lam *et al.*, 1996; King-Smith and Rose, 1997; Shen and Richards, 2012) given by

$$\Psi(x) = \gamma + \frac{1-\gamma-\lambda}{1+\exp\left(-\tilde{\beta}(x-\tilde{\alpha})\right)}, \tag{1}$$

where $\Psi(x)$ gives the proportion of words correctly identified at an SNR of x dB. The parameters $\tilde{\alpha}$ and $\tilde{\beta}$ determine the SRT and slope as discussed below while the remaining two parameters are the guess rate, γ, and the lapse rate, λ. In a context-free forced-choice experiment, γ is equal to the reciprocal of the number of alternative choices while λ is the probability of making errors at very high SNRs due either to indistinct stimuli, listener hearing impairment, or listener inattention. The slope of the PF at x may be calculated as

$$\Psi'(x) = \frac{(\Psi(x)-\gamma)(1-\lambda-\Psi(x))\tilde{\beta}}{1-\gamma-\lambda}. \tag{2}$$

A model-independent parameterization can be obtained by selecting a convenient value of ϕ (e.g., $\phi = 0.5$ or 0.75) and defining α_ϕ and β_ϕ to satisfy

[a]Electronic mail: mike.brookes@imperial.ac.uk

$$\Psi(\alpha_\phi) = \phi \quad \text{and} \quad \Psi'(\alpha_\phi) = \beta_\phi. \tag{3}$$

For any given ϕ, α_ϕ, and β_ϕ, the parameters $\tilde{\alpha}$ and $\tilde{\beta}$ in Eq. (1), respectively, can be found from the invertible relationships

$$\tilde{\beta} = \beta_\phi \frac{1-\gamma-\lambda}{(\phi-\gamma)(1-\lambda-\phi)} \tag{4}$$

and

$$\tilde{\alpha} = \alpha_\phi + \tilde{\beta}^{-1} \ln\left(\frac{1-\lambda-\phi}{\phi-\gamma}\right). \tag{5}$$

For the special case of $\phi = 0.5(1-\lambda+\gamma)$, these equations reduce to $\tilde{\alpha} = \alpha_\phi$ and $\tilde{\beta} = 4\beta_\phi(1-\gamma-\lambda)^{-1}$. The solid curve in Fig. 1 illustrates the PF generated by Eq. (1) using the parameters listed in Fig. 1. For very low and very high values of x the PF asymptotes are γ and $1-\lambda$, respectively. In this example $\phi = 0.75$ and the circle on the PF identifies the point where $\Psi(\alpha_\phi) = \phi$. The tangent to the PF at this point is shown as a dashed line and has the gradient $\Psi'(\alpha_\phi) = \beta_\phi$. For clarity, the subscripts are omitted from α_ϕ and β_ϕ in the remainder of this paper.

B. Methods for estimating the PF

A widely used procedure for measuring PFs is the method of constant stimuli (McKee *et al.*, 1985). In this approach, the PF is sampled at a number of predetermined probe levels, and multiple tests are performed at each level in a randomized order to obtain the percent-correct score at each level. The approach has the advantage that it does not need to assume a specific PF model but it is normally less efficient than methods that use a parametric model for the PF, and so it may require a large number of trials in order to obtain a good PF estimate (Leek, 2001). Notwithstanding this, the method of constant stimuli can give reliable results in a reasonable number of trials especially if prior knowledge about the underlying PF is available.

To improve the efficiency of measuring the SRT and/or slope parameters, a large number of adaptive procedures have been proposed (Hall, 1981; King-Smith *et al.*, 1994; King-Smith and Rose, 1997; Brand and Kollmeier, 2002; Shen and Richards, 2012). Among the most popular are the

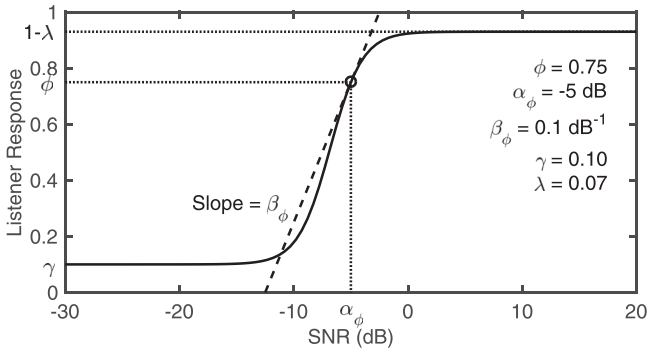


FIG. 1. Logistic PF model $\Psi(x)$ from Eq. (1) using the parameters listed in the figure. The parameters $\tilde{\alpha}$ and $\tilde{\beta}$ in Eq. (1) are determined from α_ϕ and β_ϕ using Eqs. (4) and (5).

up-down staircase procedures (Dixon and Mood, 1948; Levitt, 1971). The idea underlying these methods is to adapt the probe SNR at trial n according to the listener's response in the preceding trials so that convergence toward the SRT can be achieved rapidly. From an initial stimulus with a high expected intelligibility, the probe SNR is typically decreased or increased according to whether the listener's previous response was correct or incorrect. In order to ensure rapid convergence, the step-size governing the change in probe SNR can be adaptively selected (Kaernbach, 1991). An extension of this is the transformed up-down procedure (Levitt, 1971) in which the SNR change depends on the responses in two or more of the preceding trials. For example, with the two-down, one-up rule, the SNR is decreased after two successive correct responses, but increased after only a single incorrect response. By fitting a parametric model to the sequence of responses using a maximum likelihood criterion, it is possible to estimate both the SRT and slope at a chosen performance level. It has, however, been found that the slope estimates obtained in this way may have considerable bias (Leek *et al.*, 1992; Kaernbach, 2001; Treutwein and Hansstrasburger, 1999).

Wetherill (1963) derived expressions for the asymptotic variance of the maximum likelihood estimators of $\tilde{\alpha}$ and $\tilde{\beta}$ in Eq. (1) assuming that $\gamma = \lambda = 0$. He showed that the variance was minimized when the probe levels, x, satisfied $\Psi(x) = 0.5$ for $\tilde{\alpha}$ and $\Psi(x) \in \{0.083, 0.917\}$ for $\tilde{\beta}$. These optimum probe levels were subsequently denoted the "sweetpoints" of the PF by Laming and Marsh (1988). When wishing to estimate $\tilde{\alpha}$ and $\tilde{\beta}$ jointly, Wetherill (1963) proposed minimizing the product of their variances leading to sweetpoints satisfying $\Psi(x) \in \{0.176, 0.824\}$; he found that there was no benefit in probing at $\Psi(x) = 0.5$ in addition to these two sweetpoints. O'Regan and Humbert (1989) generalized this work and determined the sweetpoint locations for estimating $\tilde{\alpha}$ and/or $\tilde{\beta}$ when $\gamma \neq 0$ in Eq. (1).

In order to perform concurrent slope and threshold estimation, Lam *et al.* (1996) proposed an adaptive multimodal sampling scheme of the PF. Based on the analysis in Levitt (1971), four probe levels satisfying $\Psi(x) \in \{0.02, 0.16, 0.84, 0.98\}$ were sampled in each iteration. The four resultant listener responses were used to update the PF model parameters using a nonlinear fitting procedure and the sweetpoint positions were then recalculated for the next iteration.

Instead of fitting a PF model to the observed data after every new listener response, Green (1993) sampled the space of possible PF parameters onto a fixed two-dimensional grid covering 4 values of γ and 120 values of α for a total of 480 grid points. After each listener response, the likelihood of each point on the grid was updated to a new value as described in Watson and Pelli (1983). The probe level at the next trial was then determined as the α-sweetpoint of the PF corresponding to the maximum-likelihood point on the grid. A similar approach was used by King-Smith and Rose (1997) with a two-dimensional grid in $\tilde{\alpha}$ and $\log\tilde{\beta}$. Using the PF defined by the mean of the posterior distribution, the probe level at each trial was randomly selected from the two sweetpoints $\Psi(x) \in \{0.12, 0.88\}$. In Brand and Kollmeier (2002), the same technique was used in conjunction with

two randomly interleaved adaptive-staircase procedures converging to the sweetpoints $\Psi(x) \in \{0.2, 0.8\}$ that were chosen as a compromise between the α and β sweetpoints. After each trial, the maximum likelihood PF was estimated, and the SNRs corresponding to the sweetpoints recomputed. Shen and Richards (2012) proposed a similar maximum-likelihood estimation of the parameters, but included the lapse rate, λ, as an additional unknown parameter. A sweetpoint for this extra parameter was derived, and the next observation placement was computed by selecting one of the the α, λ, or β sweetpoints of the current maximum-likelihood PF estimate. Two selection schemes were evaluated; one used an adaptive procedure while the other randomly chose between the sweetpoints, and was found to converge slightly faster to the SRT estimate.

In a Bayesian adaptive procedure, the search grid is viewed as a probability distribution over the parameter space. Before the first trial, the grid is initialized with a prior distribution that encapsulates any prior knowledge about the distribution of the PF parameters. Following each trial, the posterior distribution at each point on the grid is updated using Bayes' theorem [see Eq. (6)]. At the end of the experiment, the final parameter estimates may be taken as either the mean or the mode of the posterior distribution. Rather than selecting the probe SNR to lie at an estimated sweetpoint, it is chosen in King-Smith *et al.* (1994) and Gaubitch *et al.* (2010) to minimize the expected variance of the posterior distribution at the completion of the next trial (or alternatively the next two trials). A similar approach is adopted in the Ψ procedure described in Kontsevich and Tyler (1999), but here the next probe is chosen to minimize the expected entropy of the distribution rather than its variance.

The present research extends the work on Bayesian adaptive methods and is outlined as follows. In Sec. II, the Bayesian adaptive framework for estimating PF parameters using an adaptive evaluation grid is presented. The placement of optimal degradation levels using both look one-ahead and look two-ahead methods is also derived. In Sec. III, simulations are run to test the efficiency and accuracy of the proposed estimation procedure under various test conditions. The simulation results are then discussed in Sec. IV, and conclusions are drawn in Sec. V.

II. BAYESIAN ADAPTIVE ESTIMATION

In this section, a Bayesian method for estimating the PF is presented that extends the work of Kontsevich and Tyler (1999) and Gaubitch *et al.* (2010). Section II A details the estimation of the PF parameters from their posterior distribution and describes the adaptive rescaling and resampling scheme that is used in the algorithm. Section II B describes the next probe SNR selection based on the minimization of the expected cost after one or two predicted observations.

A. Sequential Bayesian estimation

1. Estimation procedure

In the following, $\boldsymbol{\theta} = [\alpha, \log(\beta)]^T$ represents the threshold and log-slope parameters of a PF where $\log(\,)$ denotes a base-10 logarithm. Parameterizing the PF using the slope in the log-domain is convenient for several reasons: it is unbounded, its posterior distribution is found to be approximately Gaussian, and errors in the log-slope are approximately proportional to percentage errors in the slope. The parameter space is discretized with K grid points, $\boldsymbol{\theta}_k$, for $k = 1, ..., K$. The parameterized PF, $\Psi(x|\boldsymbol{\theta}_k)$, gives the probability of observing a correct response at an SNR of x dB with PF parameters $\boldsymbol{\theta}_k$. The guess rate, γ, and lapse rate, λ, are assumed to be fixed in advance; the choice of λ is discussed in Sec. II C.

At trial n, a degraded speech sample at an SNR of x_n dB is presented to a listener and $r_n = 1$ or 0 according to whether their response is correct or incorrect, respectively. The row vector $\mathbf{z}_n = [x_1, r_1, ..., x_n, r_n]$, which is always of even length, contains all the observed data up to trial n and the discretized probability distribution, $P(\boldsymbol{\theta}_k|\mathbf{z}_n)$, gives the posterior distribution of the PF parameters after n trials. After trial n, the posterior distribution $P(\boldsymbol{\theta}_k|\mathbf{z}_n)$ is updated using Bayes theorem,

$$P(\boldsymbol{\theta}_k|\mathbf{z}_n) = \frac{P(\boldsymbol{\theta}_k|\mathbf{z}_{n-1})P(r_n|x_n, \boldsymbol{\theta}_k)}{\sum_k P(\boldsymbol{\theta}_k|\mathbf{z}_{n-1})P(r_n|x_n, \boldsymbol{\theta}_k)}, \tag{6}$$

with

$$P(r_n|x_n, \boldsymbol{\theta}_k) = \begin{cases} 1 - \Psi(x_n|\boldsymbol{\theta}_k), & r_n = 0 \\ \Psi(x_n|\boldsymbol{\theta}_k), & r_n = 1 \end{cases} \tag{7}$$

and $P(\boldsymbol{\theta}_k|\mathbf{z}_0)$ as the prior distribution.

Following trial n, the mean, $\boldsymbol{\mu}(\mathbf{z}_n)$, and variance, $\mathbf{v}(\mathbf{z}_n)$, of the parameter vector $\boldsymbol{\theta} = [\alpha, \log(\beta)]^T$ are given by

$$\boldsymbol{\mu}(\mathbf{z}_n) = \sum_k \boldsymbol{\theta}_k P(\boldsymbol{\theta}_k|\mathbf{z}_n), \tag{8}$$

$$\mathbf{v}(\mathbf{z}_n) = \sum_k (\boldsymbol{\theta}_k - \boldsymbol{\mu}(\mathbf{z}_n))^2 P(\boldsymbol{\theta}_k|\mathbf{z}_n), \tag{9}$$

where, in Eq. (9), $(\,)^2$ acts elementwise on a vector.

In intelligibility tests that use sentences as tokens, either a sentence may be scored as a single unit or else each content word in the sentence may be scored independently. In the latter case, the estimated PF relates to word intelligibility. Context effects, however, mean that the constituent words in a sentence are not perceived independently. This phenomenon may be modeled by modifying the expression for $P(r_n|x_n, \boldsymbol{\theta}_k)$ in Eq. (7) to take into account the number of correct and incorrect responses in the sentence together with the j factor described in Boothroyd and Nittrouer (1988).

2. Approximation of the posterior distribution

The parameter space is sampled by a two-dimensional evaluation grid of possible values for $\boldsymbol{\theta}$. If the grid were fixed, it would need to encompass the entire range of possible parameter values and be fine enough to provide adequate precision for accurate estimation of the PF parameters. In order to avoid imposing any prior restriction on the possible

values of $\boldsymbol{\theta}$, the evaluation grid is instead adaptively rescaled and resampled as required.

To obtain an updated evaluation grid, the current estimate of the posterior distribution $P(\boldsymbol{\theta}|\mathbf{z}_n)$ is used to compute the mean, $\boldsymbol{\mu}(\mathbf{z}_n)$, and standard deviation, $\sqrt{\mathbf{v}(\mathbf{z}_n)}$, according to Eqs. (8) and (9). The rescaled evaluation grid is then defined, for each component, to cover the range $\boldsymbol{\mu}(\mathbf{z}_n)\pm 4\sqrt{\mathbf{v}(\mathbf{z}_n)}$. The choice of ± 4 standard deviations ensures (as a consequence of Chebyshev's inequality) that, regardless of the distribution's shape, at least 93.75% of its probability mass will lie within the search grid.

Whenever the grid is changed in either axis, it is necessary to interpolate the old values of $P(\boldsymbol{\theta}|\mathbf{z}_n)$ onto the new grid. The method of evaluating $P(\boldsymbol{\theta}|\mathbf{z}_n)$ for the rescaled grid depends on its relationship to the existing grid:

- If the limits of the rescaled grid match those of the existing grid within $\pm 10\%$, then no rescaling is performed. This avoids repeated interpolation by small amounts.
- If the range of the rescaled grid lies entirely within the existing grid, then the new values of $P(\boldsymbol{\theta}|\mathbf{z}_n)$ are calculated by quadratic interpolation in $\log(P(\boldsymbol{\theta}|\mathbf{z}_n))$. This form of interpolation is chosen because it has low computational complexity and is exact for a Gaussian distribution. The interpolation procedure is illustrated in Fig. 2 in which the left plot depicts $\log\left(P(\boldsymbol{\theta}|\mathbf{z}_n)\right)$ for a grid covering the range $\alpha \in [-4, 2]$ and $\ln(\beta) \in [-4, 0.5]$. In the right plot this has been interpolated in the $\log(\beta)$ direction onto a grid covering $\alpha \in [-4, 2]$ and $\ln(\beta) \in [-2.7, -0.6]$.
- If neither of the previous cases apply, $P(\boldsymbol{\theta}|\mathbf{z}_n)$ is recalculated from scratch from the values in $\mathbf{z}_n$. Although this recalculation entails additional computation, it happens relatively rarely since $\mathbf{v}(\mathbf{z}_n)$ normally decreases with n after the first few trials.

This adaptive rescaling and resampling of the evaluation grid serves two purposes. From a computational point of view, it allows us to keep the number of discretized values for $\boldsymbol{\theta}$ to a minimum because the grid contracts around the mean, $\boldsymbol{\mu}(\mathbf{z}_n)$, of the posterior distribution. This ensures that the method remains computationally tractable while still offering a fine discretization of the parameter space at convergence. The principal benefit of the technique is that it avoids the need to specify the range of possible parameter values in advance since the grid will automatically expand if required.

An example of the grid adaptation is illustrated in Fig. 3 which shows, as a function of trial number, the bounds of the evaluation grid (dashed lines) and the current SRT estimate. In this example, the true SRT is 30 dB, which lies outside the initial range for the SRT evaluation grid of ± 20 dB. The first two responses result in an SRT estimate that approaches the upper limit of the evaluation grid and this causes the grid to expand rapidly to ensure that the SRT estimate lies near the center of the grid. From trial 7 onward, the grid contracts as both the SRT estimate and the evaluation grid converge together. The process is very stable and self-correcting. If, for example, a sequence of spurious early responses were to cause the evaluation grid to converge initially around a false maximum in $P(\boldsymbol{\theta}|\mathbf{z}_n)$, subsequent responses would drive the estimated SRT to the edge of this grid, which would trigger it to expand sufficiently to include the true SRT.

B. Next probe SNR selection

In order to obtain a reliable estimate of $\boldsymbol{\theta}$ with as few trials as possible, the SNR at which to probe at the next trial must be chosen carefully so as to optimize the efficiency. Thus, the next probe SNR is selected in order to minimize a cost function, $\mathbf{c}(\mathbf{z}_n)$, that measures the uncertainty in the estimate of $\boldsymbol{\theta}$. Two alternative cost functions are considered here: (i) the variance of the $\boldsymbol{\theta}$ estimate, $\mathbf{v}(\mathbf{z}_n)$, which is computed from Eq. (9), and (ii) the differential entropy $\mathbf{h}(\mathbf{z}_n)$, which is computed as

$$\mathbf{h}(\mathbf{z}_n) = \begin{bmatrix} \Delta_\alpha \sum_k P(\alpha_k|\mathbf{z}_n)\ln(P(\alpha_k|\mathbf{z}_n)) \\ \Delta_{\log\beta} \sum_k P(\log(\beta_k)|\mathbf{z}_n)\ln(P(\log(\beta_k)|\mathbf{z}_n)) \end{bmatrix} \tag{10}$$

in which $P(\alpha_k|\mathbf{z}_n)$ and $P(\log(\beta_k)|\mathbf{z}_n)$ are obtained by marginalizing $P(\boldsymbol{\theta}_k|\mathbf{z}_n)$ over $\log(\beta_k)$ and α_k, respectively. The factors Δ_α and $\Delta_{\log\beta}$ are the evaluation grid spacings and their inclusion ensures that the value of $\mathbf{h}(\mathbf{z}_n)$ is largely unaffected when the grid is rescaled. Although Kontsevich and Tyler

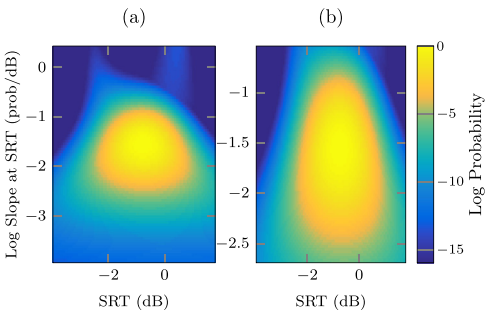


FIG. 2. (Color online) The left plot shows the posterior distribution, $P(\boldsymbol{\theta}|\mathbf{z}_n)$, evaluated on a grid covering the range $\alpha \in [-4, +2]$ and $\ln(\beta) \in [-4, +0.5]$. In the right plot, quadratic interpolation has been used to evaluate $P(\boldsymbol{\theta}|\mathbf{z}_n)$ on a new grid that has been contracted in the $\log(\beta)$ direction so that now $\ln(\beta) \in [-2.7, -0.6]$.

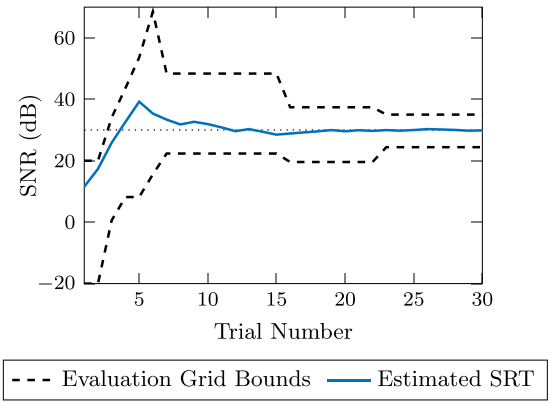


FIG. 3. (Color online) The solid line shows the estimated SRT versus trial number, n, as it converges toward $\alpha_{\text{True}} = 30$ dB. The dashed lines show the bounds of the evaluation grid which first expand from their initial values of $[-20, 20]$ dB and then contract around α_{True}.

(1999) suggest minimizing the entropy of the joint distribution, $P(\boldsymbol{\theta}_k|\mathbf{z}_n)$, this does not guarantee that the marginal distributions of the individual components have low entropy or low variance. Accordingly, the component entropies are included separately in $\mathbf{h}(\mathbf{z}_n)$ and the algorithm minimizes their weighted average as described in Eqs. (13) and (14). The entropy and the variance are both measures of how concentrated the distribution is around its mode or mean; the difference between them is that the variance penalizes large deviations from the mean more heavily.

1. Look one-ahead

A set, X, of possible probe SNRs is considered at each trial; the choice of X is discussed in Sec. II B 3. For all combinations of potential probe SNR, $x \in X$, and listener response, $r \in \{0, 1\}$, the posterior distribution of $\boldsymbol{\theta}_k$ is found by using the Bayesian update

$$P\left(\boldsymbol{\theta}_k|[\mathbf{z}_n, x, r]\right) = \frac{P(\boldsymbol{\theta}_k|\mathbf{z}_n)P(r|x, \boldsymbol{\theta}_k)}{\sum_k P(\boldsymbol{\theta}_k|\mathbf{z}_{n-1})P(r|x, \boldsymbol{\theta}_k)}. \tag{11}$$

From this, the cost function, $\mathbf{c}([\mathbf{z}_n, x, r])$ can be evaluated as either the variance, $\mathbf{v}([\mathbf{z}_n, x, r])$, from Eq. (9) or the differential entropy, $\mathbf{h}([\mathbf{z}_n, x, r])$, from Eq. (10).

The probability, $P(r|x, \mathbf{z}_n)$, of observing response r from probe SNR x is computed by marginalizing $P(r, \boldsymbol{\theta}|x, \mathbf{z}_n)$ over $\boldsymbol{\theta}$ as

$$P(r|x, \mathbf{z}_n) = \sum_k P(r|x, \boldsymbol{\theta}_k)P(\boldsymbol{\theta}_k|\mathbf{z}_n). \tag{12}$$

The probe SNR for trial $n+1$ is then chosen to minimize the weighted sum of the parameter costs in $\mathbf{c}([\mathbf{z}_n, x, r])$. That is,

$$x_{n+1} = \arg\min_{x \in X} \sum_{r=0}^{1} \mathbf{w}^T \mathbf{c}([\mathbf{z}_n, x, r])P(r|x, \mathbf{z}_n), \tag{13}$$

where

$$\mathbf{w} = [1 - \kappa, \ \kappa]^T \tag{14}$$

is a user-selected weight vector in which $\kappa \in [0, 1]$ determines the relative weights given to the costs of α and β. Note that, since the SRT and log-slope have different units, the significance of the weights depends on the units chosen, as well as on the cost function that is used. The choice of the κ value depends on the relative precision requirements of the SRT and the slope and is made available to the experimenter.

2. Look two-ahead

The method described above determines the next probe SNR, x_{n+1}, by minimizing the expected value of the cost function after the next trial. Estimating the slope, however, requires test stimuli at two different locations along the PF (Levitt, 1971; Brand and Kollmeier, 2002; Shen and Richards, 2012). Although look one-ahead search may provide such probe SNRs, it cannot do so in an informed manner as the probe SNR to improve the $\boldsymbol{\theta}$ estimate may not be the one that causes the biggest immediate decrease in the cost function. Instead, it seems reasonable to choose a pair of probe SNRs jointly to give the greatest expected improvement in the cost function after the next two trials.

Consider $(x, y) \in X \times X$ and $(r, s) \in \{0, 1\} \times \{0, 1\}$ to be the next two probe SNRs and their corresponding responses. A natural strategy would be to find the pair of probe SNRs, (x, y), that minimizes the expected cost after trial $n+2$ and to select x as the next probe. This strategy is suboptimum, however, because it does not take into account that the second probe SNR, y, can be chosen after the response, r, to the first probe is known.

Instead, $d(\mathbf{z}_n, x, r)$ is defined to be the lowest possible expected value of the cost function after trial $n+2$ given that the probe SNR and listener response at trial $n+1$ are x and r, respectively. This may be calculated as

$$d(\mathbf{z}_n, x, r) = \min_{y \in X} \sum_{s=0}^{1} \mathbf{w}^T \mathbf{c}([\mathbf{z}_n, x, r, y, s])P(s|y, [\mathbf{z}_n, x, r]), \tag{15}$$

where $\mathbf{c}(\ldots)$ is either equal to $\mathbf{v}(\ldots)$ from Eq. (9) or to $\mathbf{h}(\ldots)$ from Eq. (10) depending on the choice of cost function, and $P(s|y, [\mathbf{z}_n, x, r])$ is given by Eq. (12). The probe SNR for trial $n+1$ should minimize the expected value of this quantity and so

$$x_{n+1} = \arg\min_{x \in X} \sum_{r=0}^{1} d(\mathbf{z}_n, x, r)P(r|x, \mathbf{z}_n). \tag{16}$$

It is worth noting that, although the look two-ahead procedure considers the effect of different choices for x_{n+2} in Eq. (15), it only actually makes a decision about x_{n+1}. As will be seen in Sec. III C, the probe placements chosen by the look two-ahead method are, in practice, very similar to those chosen by the look one-ahead method. It appears that the expected reduction in the cost function is dominated by the choice made for x_{n+1} and that the influence of the speculative choice for x_{n+2} is less important.

3. Candidate probe SNRs

In some circumstances the permitted probe SNRs are predetermined (e.g., if the test stimuli have been pre-recorded). In other circumstances, the stimuli can be generated on the fly at an arbitrary SNR. In this situation, a discrete set of candidate SNRs, X, is determined and one of these is then selected as the next probe SNR. To ensure that the optimum probe SNR is included in the evaluation, the set X should be chosen to encompass the range of SNRs corresponding to performance levels between 5% and 95% of the unknown true PF.

To ensure the desired range of probe SNRs is covered, R representative PF parameters sets, $\boldsymbol{\theta}_i$ for $i \in [0, R-1]$, are placed around an elliptical iso-probability contour at a distance of two standard deviations from the mean of the posterior distribution as illustrated by the small circles in Fig. 4.

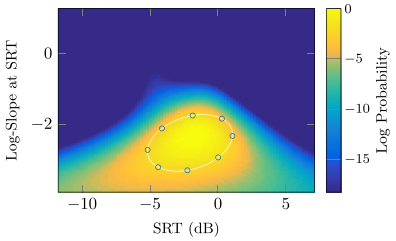


FIG. 4. (Color online) Representative PF parameters, $\boldsymbol{\theta}_i$, placed along the elliptical iso-probability contour of $P(\boldsymbol{\theta}|\mathbf{z}_n)$ at two standard deviations from the mean.

In this work, the value $R = 8$ is used to ensure that the contour is adequately sampled.

Using the eigen-decomposition of the covariance matrix of the posterior distribution with

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\theta}_k|\mathbf{z}_n) &= \sum_k \left[(\boldsymbol{\theta}_k - \boldsymbol{\mu}(\mathbf{z}_n))(\boldsymbol{\theta}_k - \boldsymbol{\mu}(\mathbf{z}_n))^T\right] P(\boldsymbol{\theta}_k|\mathbf{z}_n) \\ &= \mathbf{QDQ}^T, \end{aligned} \tag{17}$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{Q}$ is orthogonal, then $\boldsymbol{\theta}_i$ can be written as

$$\boldsymbol{\theta}_i^T = \boldsymbol{\mu}(\mathbf{z}_n)^T + 2\left[\cos\left(\frac{\pi}{4}.i\right), \sin\left(\frac{\pi}{4}.i\right)\right]\mathbf{Q}\sqrt{\mathbf{D}}\mathbf{Q}^T, \tag{18}$$

where $\sqrt{.}$ acts elementwise on a matrix and $i = \{0, .., 7\}$.

For each of the R representative PFs defined by the $\boldsymbol{\theta}_i$, the SNRs corresponding to the 5% and 95% correct scores are computed and the two extreme SNR values from this set are used to define the range that needs to be covered by the list of candidate probe SNRs. The candidate probe SNRs are then linearly spaced in dB within this range. Any external restrictions on the range of permitted SNRs can be imposed by the experimenter at this stage.

C. Lapses of attention—Handling outliers

In the UML algorithm of Shen and Richards (2012), the lapse rate λ is optionally regarded as an unknown parameter of the PF that can be estimated alongside α and β. This capability can be especially useful as a diagnosis tool to estimate the intelligibility in quiet for hearing-impaired listeners. However, in a psychoacoustic experiment with normal-hearing listeners, we believe that the value of λ is likely to vary erratically over time. Accordingly, instead of considering λ to be a free parameter, it is given a fixed small value (e.g., 2%). Based on the evaluations in Sec. III, the precise choice of λ is unimportant unless it is very much less than the true lapse rate. To accommodate the possibility of a listener having an unusually high rate of attention lapses, the experimenter is given the option of re-computing the posterior pdf of $P(\boldsymbol{\theta}|\mathbf{z}_n)$ from the saved experimental results while excluding outliers. This means that from the current estimate of $P(\boldsymbol{\theta}|\mathbf{z}_n)$, incorrect responses at probe SNRs for which the PF exceeds, for example, 0.95 are marked as outliers and ignored when re-computing the posterior density.

III. SIMULATION RESULTS

Two sets of Monte Carlo simulations were run in order to evaluate the proposed Bayesian framework under various test conditions and to compare it to competing methods. The first set evaluated six specific experimental setups and examined the convergence of the estimated PF parameters as a function of trial number, n. The second set of simulations fixed the number of trials at a realistic value of $n = 50$ and investigated the effect on performance of systematically varying the parameters of the true PF of a virtual listener.

Two competing methods were compared with four versions of the proposed estimation algorithm making six algorithms in all

(1) The adaptive maximum-likelihood (AML) technique presented in Brand and Kollmeier (2002) using the parameter values given for the "adaptive procedure A2." The adaptive level placement used in this method, described briefly in Sec. I B, is based on two interleaved staircase procedures that target a pair of sweetpoints, $\Psi(x) \in \{0.2, 0.8\}$, independently of the estimated PF. The authors adapted their procedure to the sentence scoring used in their evaluation example and so it has a somewhat different goal from the other algorithms presented in this study. The method was modified to include a prior distribution and used a fixed evaluation grid of 150×150 linearly spaced values for $\boldsymbol{\theta}$ parameterized by SRT and log-slope.

(2) The updated maximum-likelihood (UML) procedure presented in Shen and Richards (2012) with two-down, one-up stimulus placement strategy. The implementation used was taken from V2.2 of the toolbox presented in Shen *et al.* (2015), with the λ parameter fixed at λ_{Alg} to ensure a fair comparison with other algorithms. The fixed evaluation grid was defined by 81 logarithmically spaced values for the slope and 81 linearly spaced SRT values. Note that this algorithm uses $\tilde{\alpha}$ and $\tilde{\beta}$ from Eq. (1) as internal parameters rather than α and β.

(3) The proposed Bayesian adaptive procedure using either look one-ahead (L1A) or look two-ahead (L2A) for the next probe SNR selection and using either an entropy-based cost function (L1A-Ent and L2A-Ent) or a variance-based cost function (L1A-Var and L2A-Var). The adaptive evaluation grid comprised 40 linearly spaced values for the SRT and 21 linearly spaced values for the log-slope. At each trial, 30 candidate probe SNRs were considered. The default value of the weighting factor κ from Eq. (14) was 0.5. No rejection of outliers was used when computing $\boldsymbol{\mu}(\mathbf{z}_n)$. The MATLAB implementation associated with this paper is freely available as the *psycest* function from the Voicebox Toolbox (Brookes, 2016).

All methods were initialized with probe SNRs in the range $(-20, 20)$ dB, and evaluation grids covering the range $(-20, 20)$ dB for α and $(0.01, 0.5)$ dB^{-1} for β on a logarithmic scale. A Gaussian prior $P(\boldsymbol{\theta}|\mathbf{z}_0)$ was used, with a mean at the center of the evaluation grid and the standard deviation for each variable set to half the range of the grid. The

algorithms used a Logistic PF model, M_{Alg}, with fixed parameters $\gamma_{\text{Alg}} = 0.01$ and $\lambda_{\text{Alg}} = 0.02$.

For each experimental setup and each estimation algorithm, $N = 1000$ Monte Carlo simulations were run where each simulation comprised a sequence of 300 trials. At each trial, the algorithm under test selects a probe SNR and receives a binary-valued response from a virtual listener. The virtual listener's response is generated according to a "true" PF that is defined by a set of five parameters: α_{True}, β_{True}, γ_{True}, λ_{True}, and M_{True}. The PF model, M_{True}, was either a Logistic function [following Eq. (1)] or a cumulative Gaussian (Levitt, 1971; Hall, 1981). In order to assess performance, the estimation errors were computed as

$$e_\alpha(n) = \hat{\alpha}_n - \alpha_{\text{True}} \quad \text{and} \quad e_\beta(n) = \log\left(\frac{\hat{\beta}_n}{\beta_{\text{True}}}\right) \tag{19}$$

where $\hat{\alpha}_n$ and $\hat{\beta}_n$ are the estimated parameters after n trials obtained from $\boldsymbol{\mu}(\mathbf{z}_n)$ in Eq. (8). Values of e_β equal to ± 0.1 correspond to percentage errors of $+26\%$ and -20%, respectively, in the estimation of the slope. The root-mean-squared error (RMSE) and bias of parameter $* \in \{\alpha, \beta\}$ are computed as

$$\text{RMSE}_*(n) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_*^2(n, i)} \tag{20}$$

and

$$\text{BIAS}_*(n) = \frac{1}{N}\sum_{i=1}^{N} e_*(n, i), \tag{21}$$

where $e_*(n, i) = e_\alpha(n)$ or $e_\beta(n)$ for simulation run i.

A. Convergence evaluations

The first set of experiments uses the six experimental setups shown in Table I whose true PFs are plotted in Fig. 5. For each setup, A–F, Table I lists the model parameters of the true PF (subscripted "True") and the model parameters assumed by the algorithms (subscripted "Alg"). Setup A represents typical parameter values for the PF: at $\phi = 0.5$, the SRT is $\alpha_{\text{True}} = -5$ dB and the slope is $\beta_{\text{True}} = 0.075$ dB^{-1}, the mean slope observed in MacPherson and Akeroyd (2014). The guess rate, $\gamma_{\text{True}} = 1\%$, and lapse rate, $\lambda_{\text{True}} = 2\%$, correspond to low token predictability and an attentive listener. In each of the remaining five setups, B–F, one of the model parameters is changed substantially while the others retain their values from setup A.

TABLE I. True PF parameters and algorithm PF parameters used in the six experimental setups.

Setup	ϕ	α_{True} (dB)	β_{True} (dB^{-1})	γ_{True}	λ_{True}	M_{True}	γ_{Alg}	λ_{Alg}	M_{Alg}
A	50%	−5	0.075	1%	2%	Logistic	1%	2%	Logistic
B	50%	−5	0.075	1%	2%	Cumulative Gaussian	1%	2%	Logistic
C	50%	25	0.075	1%	2%	Logistic	1%	2%	Logistic
D	50%	−5	0.020	1%	2%	Logistic	1%	2%	Logistic
E	50%	−5	0.075	1%	10%	Logistic	1%	2%	Logistic
F	75%	−5	0.075	1%	2%	Logistic	1%	2%	Logistic

The convergence curves for setups A–F are shown in Fig. 6. Each column presents the results for one of the experimental setups and, in each case, the upper two plots give the RMSE of the SRT and slope, while the lower two plots give the corresponding mean bias. All quantities are plotted against the trial number, n, on a logarithmic scale where $5 \le n \le 300$. Each plot includes curves for four algorithms: (i) AML, Brand and Kollmeier (2002), (ii) UML, Shen and Richards (2012), (iii) L1A-Ent, and (iv) L2A-Ent. The latter two are the proposed look one-ahead and two-ahead algorithms, respectively, using an entropy cost function. For clarity, L1A-Var and L2A-Var are not shown in these plots. The four evaluated algorithms are distinguished by line color and/or line style according to the legend above Fig. 6. The true PF parameter values for setup A are shown in the top left plot; the other plots in the top row show only the parameter values that differ from setup A. For all plots $\gamma_{\text{Alg}} = 0.01$, $\lambda_{\text{Alg}} = 0.02$, and $M_{\text{Alg}} =$ Logistic.

The performance curves for setup A are plotted in the leftmost column. The upper plot in this column, demonstrates that all four algorithms have very similar convergence curves for SRT RMSE. The L1A-Ent and L2A-Ent algorithms converge slightly faster than AML and UML; the horizontal spacing between L1A-Ent and AML corresponds to a factor of up to 1.5 in n. The second plot in the column shows that for $n > 70$, the log-slope RMSE is very similar for all algorithms but that for low n the UML algorithm has the lowest RMSE. The lower two plots in this column show that for large n, all algorithms give unbiased estimates of both SRT and log-slope. The UML algorithm underestimates both the SRT and the slope for $n < 70$, whereas the other algorithms have near-zero SRT bias but overestimate the slope.

In setup B the true PF follows a cumulative Gaussian, whereas the estimation algorithms assume a Logistic function. The small difference between these is visible in Fig. 5 at an SNR of around 5 dB. From the similarity between the first two columns of Fig. 6, the model mismatch has little effect on any of the plots. The only noticeable difference is that the UML algorithm converges slightly slower in SRT and slightly faster in log-slope.

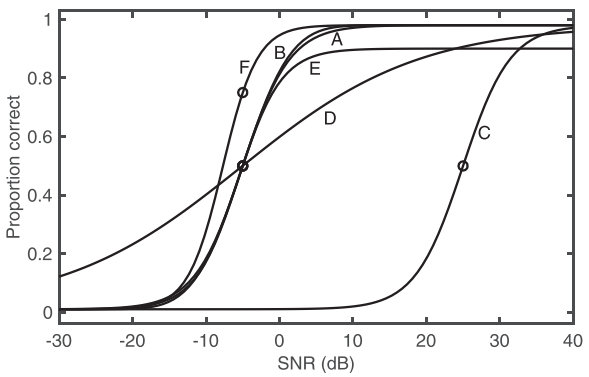


FIG. 5. The PFs of the six setups listed in Table I. In each case, the circle indicates the target SRT for the estimation algorithms.

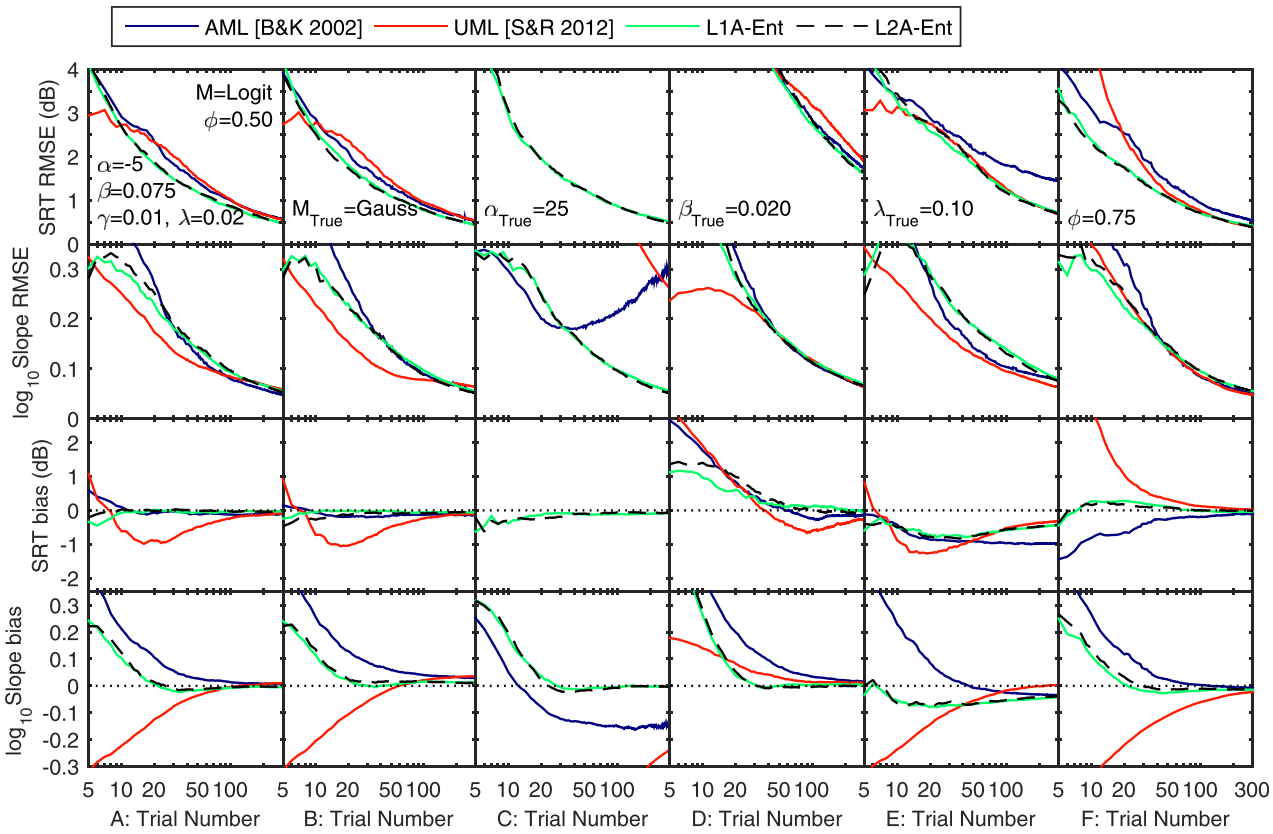


FIG. 6. (Color online) Convergence evaluations**:** Each column shows performance curves for one of the six experimental setups listed in Table I. The upper two plots in each column show the variation versus trial number, $n \in [5, 300]$, of the RMSE in the estimated SRT, $\hat{\alpha}$, and log-slope, $\log \hat{\beta}$. The lower two plots show the mean bias in these quantities. The true PF parameter values for setup A are shown in the top left plot; the other plots in the top row show only the parameter values that differ from setup A. For all plots $\gamma_{\text{Alg}} = 0.01$, $\lambda_{\text{Alg}} = 0.02$, and $M_{Alg} = \text{Logistic}$. Curves are plotted for four algorithms: (i) AML from Brand and Kollmeier (2002), (ii) UML from Shen and Richards (2012), (iii) L1A-Ent, and (iv) L2A-Ent. The latter two are the proposed look one-ahead and two-ahead algorithms, respectively, using an entropy cost function.

In setup C, shown in column three of Fig. 6, the true SRT is +25 dB and, because this value lies outside the search grid of the AML and UML algorithms, they are unable to converge. These two algorithms are omitted from plots one and three of the column because their errors exceed the plotted range. In contrast, the adaptive search grid of the L1A-Ent and L2A-Ent algorithms means that their convergence curves are almost the same as for setup A but with a very slightly delayed convergence.

In setup D, the true PF has a shallow slope of $\beta_{\text{True}} = 0.02$ dB^{-1}. As is clear from the top plot in column four of Fig. 6, the effect of this is that all four algorithms converge more slowly both in SRT (by a factor of 5 in n) and in log-slope (by a factor of 1.2 in n). The UML algorithm again converges fastest in log-slope but only for $n < 30$.

In setup E, shown in the fifth column, the lapse rate is increased to the high value of $\lambda_{\text{True}} = 10\%$. All algorithms converge somewhat more slowly than for setup A; the AML algorithm is most affected for SRT RMSE and the L1A-Ent and L2A-Ent algorithms are most affected for slope RMSE. As can be seen from the lower two plots, all four algorithms now have biased SRT estimates and all but UML have biased log-slope estimates. At $n = 300$, the L1A-Ent and L2A-Ent algorithms have an SRT bias of −0.4 dB and a log-slope bias of −0.04, which corresponds to a 8% underestimate of the slope. The UML algorithm has the lowest bias: −0.25 dB in SRT and +0.01 in log-slope, which corresponds to +2% in slope.

Finally, in setup F, the target SRT is changed from a level of 50% to a level of 75%, which, for speech intelligibility, is a more practically relevant threshold. The plots in the rightmost column show that the L1A-Ent and L2A-Ent algorithms are affected very little but that the AML and UML algorithms converge somewhat more slowly for $n < 30$ in both SRT and log-slope.

B. Parameter sensitivity

The second set of experiments examines how the performance of each algorithm depends on the true PF parameters and on the weight, κ, from Eq. (14) when the number of trials is fixed at $n = 50$. Each column of Fig. 7 shows the effect of varying one parameter while the remaining parameters are fixed at the values from setup A in Table I. In each column the vertical dotted line corresponds exactly to setup A and, therefore, to the results in the leftmost column of Fig. 6 at position $n = 50$. Curves are plotted for six algorithms: (i) AML from Brand and Kollmeier (2002), (ii) UML from Shen and Richards (2012), (iii) L1A-Ent, (iv) L2A-Ent, (v) L1A-Var, and (vi) L2A-Var. The variance-based algorithms, L1A-Var and L2A-Var have been omitted from columns 1, 3, and 4 for clarity. The AML and UML algorithms are omitted from column 5 since they do not use the κ parameter.

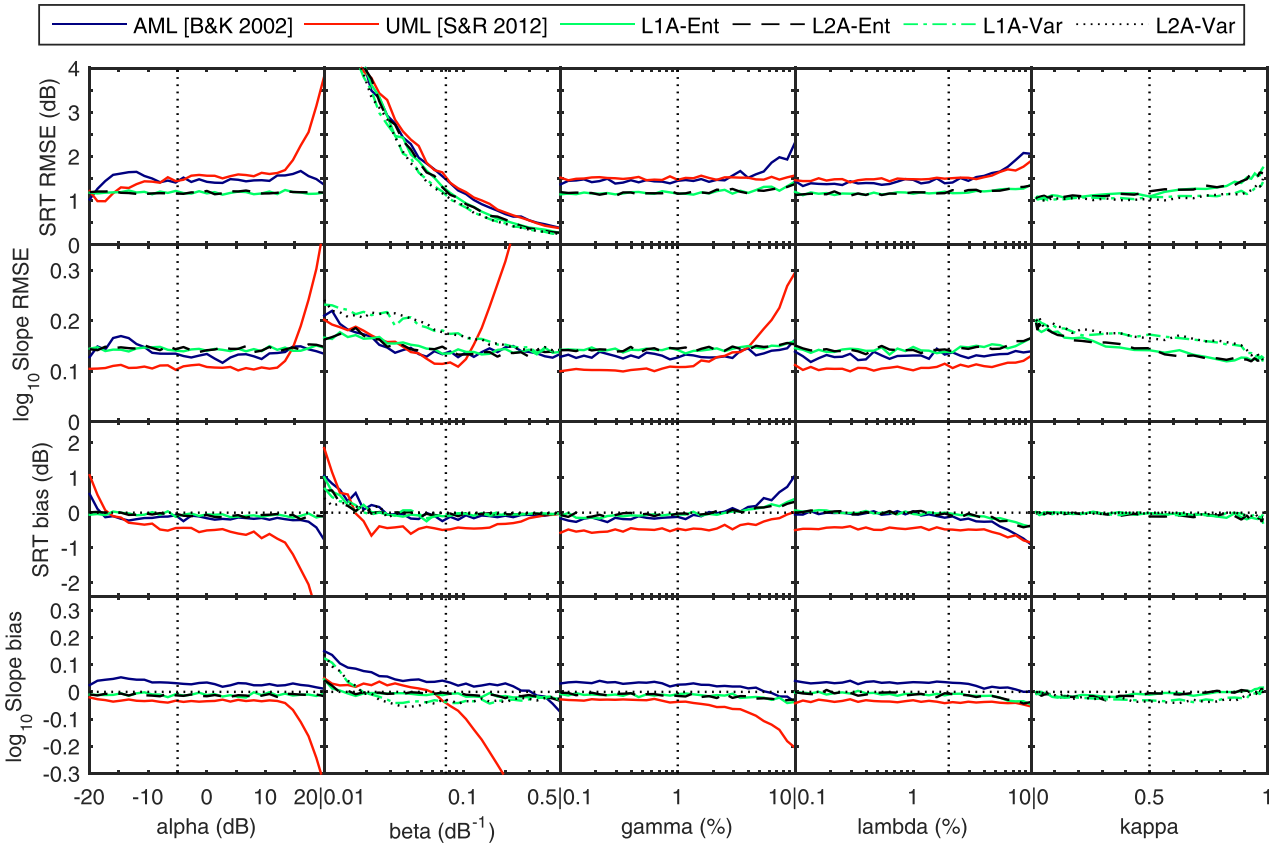


FIG. 7. (Color online) Parameter sensitivity: Each of the first four columns shows how the performance of the algorithms changes as one parameter of the true PF is varied with the other parameters held at the values from setup A (indicated by the vertical dotted lines). The final column shows how performance varies as a function of the weight, κ, in Eq. (14). In all cases the number of trials is fixed at $n = 50$. Within each column, the upper two plots show the RMSE in the estimated SRT, $\hat{\alpha}$, and log-slope, $\log \hat{\beta}$, and the lower two plots show the mean bias in these quantities. Curves are plotted for six algorithms: (i) AML from Brand and Kollmeier (2002), (ii) UML from Shen and Richards (2012), (iii) L1A-Ent, (iv) L2A-Ent, (v) L1A-Var, and (vi) L2A-Var. The variance-based algorithms, L1A-Var and L2A-Var, have been omitted from columns 1, 3, and 4 for clarity. The AML and UML algorithms are omitted from column 5 since they do not use the κ parameter.

The leftmost column of Fig. 7 shows the effect of varying α_{True} over the range ± 20 dB, which corresponds to the range of the default search grid. The upper two plots show that, except for the UML algorithm, this has little effect on the RMSE of the SRT and the log-slope. For the UML algorithm, however, the RMSE of both SRT and log-slope rise rapidly when α_{True} approaches within 5 dB of the upper edge of the search grid. The lower two plots in this column show that this is associated with a rapidly increasing negative bias in both SRT and log-slope.

The second column of Fig. 7 shows the effect of varying β_{True} over the range 0.01–0.5 dB^{-1}; this encompasses the range of slopes found by MacPherson and Akeroyd (2014). The upper plot confirms the earlier finding for setup D that all the algorithms converge to the correct SRT much more slowly when β_{True} is low. The second plot shows that this is also true for the log-slope although the effect is less pronounced than for the SRT. The variance-based algorithms, L1A-Var and L2A-Var, are more affected by low values of β_{True} than the other algorithms, which are very similar. Uniquely, the slope estimate of the UML algorithm is also affected adversely by high values of $\beta_{\text{True}} > 0.15$, which result in a large negative bias and high RMSE.

The third and fourth columns of Fig. 7 show that γ_{True} and λ_{True} have relatively little effect on the performance of any of the algorithms over the range 0.1%–10%. At high values of γ_{True}, the SRT estimate of the AML algorithm and the slope estimate of the UML algorithm become increasingly biased with accompanying increases in RMSE. At high values of λ_{True}, all algorithms have a negative SRT bias and an associated increase in RMSE although the L1A-Ent and L2A-Ent algorithms are affected less than AML and UML. The maximum plotted value of $\lambda = 10\%$ corresponds to the point $n = 50$ in the fifth column of Fig. 6.

Finally, the fifth column shows the effect of varying the weight, κ from Eq. (14), which controls the relative importance of SRT errors and slope errors in the cost function of the proposed algorithms. Varying κ has almost no effect on the bias of the SRT or the log-slope but, as would be expected, the SRT RMSE error increases for high κ and the slope RMSE increases for low κ because the relative rate of convergence is affected. The effect on the SRT RMSE is lowest for L1A-Var and L2A-Var and the effect on the slope RMSE is lowest for L1A-Ent and L2A-Ent. From these curves, there seems to be little value in changing κ from its default value of 0.5 and it should certainly not be lower than this.

C. Probe placement

Figure 8 shows the histogram of probe SNRs for each of the 6 algorithms in all N 300-trial runs of setup A from

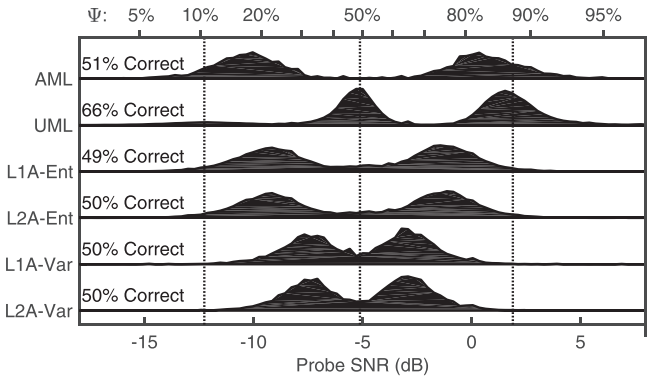


FIG. 8. For each of six algorithms, the histogram shows the distribution of probe SNR placements in $N = 1000$ runs of 300 trials for setup A from Table I. The horizontal axis shows the probe SNR on the bottom and the corresponding value of the PF along the top. The vertical dotted lines indicate the sweetpoints of this PF at SNR $= \{-12.3, -5.1, +1.9\}$ dB corresponding to $\Psi = \{0.104, 0.491, 0.879\}$.

Table I. The vertical dotted lines show the sweetpoints (Shen and Richards, 2012) for these PF parameters and the overall percentage of correct responses is also shown for each algorithm. The horizontal axis indicates the probe SNR in dB, and the corresponding PF probability is marked along the top of the graph. The AML algorithm targets the 20% and 80% points of the PF as a compromise between the SRT and slope estimation sweetpoints. The UML algorithm targets all three sweetpoints, which are recalculated after each listener response from the current maximum likelihood estimate of the PF. The values of Ψ at these sweetpoints combined with the use of a two-down, one-up selection rule results in a 66% correct-response rate. The algorithm is biased toward the upper two sweetpoints so much that the broad peak centered on the lowest sweetpoint is barely visible in the histogram. It seems reasonable to suppose that this accounts for the negative SRT bias visible in the third row of Fig. 6 for $n < 100$.

The L1A-Ent and L2A-Ent algorithms result in very similar histograms that peak at $\Psi \approx \{22\%, 76\%\}$ while the variance-based algorithms, L1A-Var and L2A-Var, have peaks that are even closer to the SRT at $\Psi \approx \{33\%, 65\%\}$. For all four algorithms, the positions of the peaks are affected by the choice of κ: the peaks lie on the two outer sweetpoints when $\kappa = 1$ but as κ is reduced toward 0, the peaks move closer together until they coalesce into a single peak at the central sweetpoint. The variance-based algorithms have narrower peaks than the entropy-based algorithms and, because of the scaling effects discussed in Sec. II B 1, lie closer to the central sweetpoint for any given value of κ. This difference explains why, in Fig. 7, the variance-based algorithms have higher RMS errors in log-slope but lower RMS errors in SRT. All four algorithms, but especially the variance-based ones, select probe SNRs in the interval between the two histogram peaks more frequently than is the case for the AML and UML algorithms. There are no significant differences in probe SNR placement between the look one-ahead and look two-ahead algorithms and this is reflected in their almost identical performance curves. Because the probe placement in the proposed algorithms is concentrated closer to the SRT, it might be expected that the algorithms would have reduced sensitivity to model mismatch when it is more severe than in setup B. This hypothesis was not investigated in the current study.

IV. DISCUSSION

Under most circumstances, all the evaluated algorithms converge to unbiased estimates of both SRT and log-slope for large n. The only exception to this noted in the simulations was when λ_{True} was much larger than the value assumed by the estimation algorithms; under this circumstance, all algorithms showed a negative bias in SRT and all but UML showed a bias in log-slope. We expect that a similar effect would result from a mismatch in γ_{True} but we did not include this in the tests since the correct value of γ_{True} is normally known. In all tests with a performance target of $\phi = 0.5$, the L1A-Ent and L2A-Ent algorithms converged fastest in SRT, whereas the UML algorithm converged fastest in slope. When the performance target was changed to $\phi = 0.75$, the convergence rates of AML and UML degraded and the L1A-Ent and L2A-Ent algorithms converged fastest in both parameters.

When the number of trials was restricted to $n = 50$, we observed some consistent differences in performance. The L1A-Ent and L2A-Ent algorithms almost always had a lower SRT RMSE than AML and UML: typically by ~ 0.4 dB. Conversely, the UML algorithm usually had a lower log-slope RMSE than the other algorithms although this advantage disappeared for β_{True} outside the range $[0.05, 0.11]$.

Because they use a fixed search grid, the AML and UML algorithms cannot converge if the SRT or slope lie outside this predefined range; for these algorithms the initial search grid should therefore be chosen to be large enough to encompass all possible outcomes. In contrast, the adaptive search grid of the proposed algorithm means that it will converge to the correct values of SRT and slope even if they lie outside the initial search range. The UML algorithm gave poor slope estimates when the true SRT, α_{True}, or true slope, β_{True}, lay close to the upper limit of the search grid.

The UML (Shen and Richards, 2012) and Ψ (Kontsevich and Tyler, 1999) procedures have been compared in the literature, and results generally show similar performance for both SRT and slope estimation (Shen, 2013; Shen *et al.*, 2015; Hatzfeld *et al.*, 2016). The Ψ procedure corresponds closely to the look one-ahead entropy-based version of the proposed method (L1A-Ent). The results obtained from our simulations broadly agree with those reported by Hatzfeld *et al.* (2016) who found, like us, that the UML algorithm was rather less sensitive than Ψ to high values of λ_{True} but recommended that $n > 80$ trials were needed to maximize the precision of the SRT estimate from the UML algorithm. Since all the algorithms use an essentially identical technique to estimate $P(\boldsymbol{\theta}_k|\mathbf{z}_n)$, differences in their performance are primarily a consequence of differences in the probe SNR placements whose histograms are shown in Fig. 8. Given the substantial differences in these histograms, it is perhaps surprising how similar the convergence curves of the six algorithms are.

The simulation results indicate that performance differences between the look one-ahead versions (L1A-Ent and L1A-Var) of the proposed algorithm and the look two-ahead versions (L2A-Ent and L2A-Var) are, in most cases, very small. This is reflected in the near-identical histograms in Fig. 8 and is consistent with the finding of King-Smith *et al.* (1994) who recommended on the grounds of computational complexity that a look one-ahead procedure should be used except for the case of two-alternative forced-choice (2AFC) experiments where the performance difference was significant. We found the lack of a performance difference surprising because we expected the look two-ahead procedures to make probe SNR choices that would result in faster slope convergence. It may be, as King-Smith *et al.* (1994) suggests, that there are some situations when this is true, but we have not discovered them. Although the look two-ahead procedure entails more computation, this is not a significant issue with current computers, which can easily perform the necessary computation in real time.

If κ is adjusted to account for the scaling effects discussed in Sec. II B 1, the differences between the entropy-based and variance-based cost functions are negligible for large n. For small n, the entropy-based cost function results in slightly better estimates of the log-slope but slightly worse estimates of the SRT. Overall, we prefer the entropy-based versions since their estimates of log-slope are more robust especially when β_{True} is small.

From the bottom two rows of Fig. 7, we see that the proposed algorithm retains a very low bias in both SRT and log-slope for wide variations in the listener PF. The bottom 2 rows of Fig. 6 indicate that, for most setups, both the SRT and log-slope bias reach zero after about 20 trials.

V. SUMMARY AND CONCLUSIONS

In an extension to previous work on the Bayesian formulation of the PF estimation problem, a new algorithm has been presented that uses an adaptive evaluation grid for the PF parameters and optimal presentation level selection using either a look one-ahead or look two-ahead procedure. The procedure converges slightly faster than competing algorithms in SRT estimation but slightly slower in log-slope estimation and both estimates have a very low bias. The procedure differs from previously proposed algorithms in two respects: (i) it does not require the range of possible PF parameters to be specified in advance and (ii) the sequence of probe SNRs optimizes the SRT and log-slope estimates at a performance level, ϕ, that can be chosen by the experimenter rather than being fixed at $\phi = 0.5$. The proposed procedure is robust to variations in the listener PF and, over a wide range of listener PF parameters, the RMS errors were 1.2 dB in SRT and 0.14 in log-slope for a 50-trial measurement. It was found that the performance differences between the look one-ahead and look two-ahead methods were negligible and that an entropy-based criterion for selecting the next stimulus was preferred to a variance-based criterion.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the EU 7th Framework Programme (FP7/2007-2013) under Grant No. ITN-GA-2012-316969, and from the Engineering and Physical Sciences Research Council under Grant No. EP/M026698/1. Data associated with this paper are available online.[1] The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

[1]http://www.commsp.ee.ic.ac.uk/~sap/projects/psycest/ (Last viewed March 30, 2017).

Boothroyd, A., and Nittrouer, S. (**1988**). "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. **84**, 101–114.
Brand, T., and Kollmeier, B. (**2002**). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am. **111**, 2801–2810.
Brookes, D. M. (**2016**). "VOICEBOX: A speech processing toolbox for MATLAB," available at http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (Last viewed March 30, 2017).
Dixon, W. J., and Mood, A. M. (**1948**). "A method for obtaining and analyzing sensitivity data," J. Am. Stat. Assoc. **43**, 109–126.
Gaubitch, N. D., Brookes, D. M., Naylor, P. A., and Sharma, D. (**2010**). "Bayesian adaptive method for estimating speech intelligibility in noise," in *Proc. AES 39th Int. Conf. on Audio Forensics*, Hillerød, Denmark, pp. 169–174.
Green, D. M. (**1993**). "A maximum-likelihood method for estimating thresholds in a yes–no task," J. Acoust. Soc. Am. **93**, 2096–2105.
Hall, J. L. (**1981**). "Hybrid adaptive procedure for estimation of psychometric functions," J. Acoust. Soc. Am. **69**, 1763–1769.
Hatzfeld, C., Hoang, V. Q., and Kupnik, M. (**2016**). "It's all about the subject-options to improve psychometric procedure performance," in *Proc. International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*, London, pp. 394–403.
Kaernbach, C. (**1991**). "Simple adaptive testing with the weighted up-down method," Percept. Psychophys. **49**, 227–229.
Kaernbach, C. (**2001**). "Slope bias of psychometric functions derived from adaptive data," Percept. Psychophys. **63**, 1389–1398.
King-Smith, P. E., Grigsby, S., Vingrys, A., Benes, S., and Supowit, A. (**1994**). "Efficient and unbiased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation and practical implementation," Vision Res. **34**, 885–912.
King-Smith, P. E., and Rose, D. (**1997**). "Principles of an adaptive method for measuring the slope of the psychometric function," Vision Res. **37**, 1595–1604.
Kontsevich, L. L., and Tyler, C. W. (**1999**). "Bayesian adaptive estimation of psychometric slope and threshold," Vision Res. **39**, 2729–2737.
Lam, C. F., Mills, J. H., and Dubno, J. R. (**1996**). "Placement of observations for the efficient estimation of a psychometric function," J. Acoust. Soc. Am. **99**, 3689–3693.
Laming, D., and Marsh, D. (**1988**). "Some performance tests of quest on measurements of vibrotactile thresholds," Percept. Psychophys. **44**, 99–107.
Leek, M. R. (**2001**). "Adaptive procedures in psychophysical research," Percept. Psychophys. **63**, 1279–1292.
Leek, M. R., Hanna, T. E., and Marshall, L. (**1992**). "Estimation of psychometric functions from adaptive tracking procedures," Percept. Psychophys. **51**, 247–256.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**(2), 467–477.
MacPherson, A., and Akeroyd, M. A. (**2014**). "Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey," Trends Hear. **18**, 1–26.
McKee, S. P., Klein, S. A., and Teller, D. Y. (**1985**). "Statistical properties of forced-choice psychometric functions: Implications of probit analysis," Percept. Psychophys. **37**, 286–298.
Miller, G., Heise, H., and Lichten, W. (**1951**). "The intelligibility of speech as a function of the context of the test materials," J. Exp. Psychol. **41**, 329–335.
Neuman, A. C., Wroblewski, M., Hajicek, J., and Rubinstein, A. (**2010**). "Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults," Ear Hear. **31**, 336–344.
O'Regan, J. K., and Humbert, R. (**1989**). "Estimating psychometric functions in forced-choice situations: Significant biases found in threshold and

slope estimations when small samples are used," Percept. Psychophys. **46**, 434–442.
Plomp, R., and Mimpen, A. M. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Int. J. Audiol. **18**, 43–52.
Shen, Y. (**2013**). "Comparing adaptive procedures for estimating the psychometric function for an auditory gap detection task," Atten. Percept. Psychophys. **75**, 771–780.
Shen, Y., Dai, W., and Richards, V. M. (**2015**). "A MATLAB toolbox for the efficient estimation of the psychometric function using the updated maximum-likelihood adaptive procedure," Behav. Res. Methods **47**, 13–26.
Shen, Y., and Richards, V. M. (**2012**). "A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention," J. Acoust. Soc. Am. **132**, 957–967.
Treutwein, B., and Hansstrasburger (**1999**). "Fitting the psychometric function," Percept. Psychophys. **61**, 87–106.
Watson, A. B., and Pelli, D. G. (**1983**). "Quest: A Bayesian adaptive psychometric method," Percept. Psychophys. **33**, 113–120.
Wetherill, G. (**1963**). "Sequential estimation of quantal response curves," J. R. Stat. Soc. Ser. B (Methodological) **25**(1), 1–48, http://www.jstor.org/stable/2984541 (Last viewed April 3, 2017).
Wichmann, F., and Hill, N. (**2001**). "The psychometric function: I. Fitting, sampling, and goodness of fit," Percept. Psychophys. **63**, 1293–1313.
Wilson, R. H., McArdle, R., and Smith, S. L. (**2007**). "An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss," J. Speech Lang. Hear. Res. **50**, 844–856.