Feasibility of interleaved Bayesian adaptive procedures in estimating the equal-loudness contour

Yi Shen,[1,a)] Celia Zhang,[2] and Zhuohuang Zhang[1]
[1]*Department of Speech and Hearing Sciences, Indiana University Bloomington, 200 S Jordan Avenue, Bloomington, Indiana 47405, USA*
[2]*Center for Hearing and Deafness, Department of Communicative Disorders and Sciences, State University of New York at Buffalo, 137 Cary Hall, Buffalo, New York 14214, USA*

(Received 8 April 2018; revised 2 October 2018; accepted 4 October 2018; published online 24 October 2018)

A Bayesian adaptive procedure, the interleaved-equal-loudness contour (IELC) procedure, was developed to improve the efficiency in estimating the equal-loudness contour. Experiment 1 evaluated the test-retest reliability of the IELC procedure using six naive normal-hearing listeners. Two IELC runs of 200 trials were conducted and excellent test-retest reliability was found at both the group and individual levels. Using the same group of listeners, Experiment 2 compared the IELC procedure to two other procedures that required frequency-by-frequency testing. One of these procedures was the commonly adopted interleaved staircase (ISC) procedure from Jesteadt [(1980). Atten. Percept. Psychophys. **28**, 85–88]. The other procedure, the interleaved maximum-likelihood (IML) procedure, was a modification of the updated maximum-likelihood procedure [Shen and Richards (2012). J. Acoust. Soc. Am. **132**, 957–967]. For each of the ISC and IML procedures, two runs of approximately 500 trials were conducted, followed by one additional IELC run. The test-retest reliability of the IELC procedure was comparable to that of the ISC and IML procedure. The accuracies of all three procedures measured in Experiment 2 were similar, which was superior to the accuracies of the IELC runs from Experiment 1, indicating a potential training effect.
 https://doi.org/10.1121/1.5064790

[GCS] Pages: 2363–2374

I. INTRODUCTION

Despite the fact that loudness is one of the most fundamental aspects of auditory perception, the quantification of loudness through behavioral experiments can be difficult and time-consuming (e.g., Marks and Florentine, 2011). A common experimental paradigm involves loudness comparison between pairs of stimuli, from which the level of a test stimulus required to match the loudness of a standard stimulus, i.e., the point of subjective equality (PSE), can be estimated. A number of potential sources of biases can influence the validity of PSE measurements, among which (1) stimulus range and (2) stimulus sequential order are known to bias loudness comparison results (e.g., Marks, 1988, 1993).

In an effort to address these issues, Jesteadt (1980) proposed an adaptive procedure for estimating the PSE. In this procedure, two interleaved adaptive tracks are measured simultaneously. In one track (i.e., the high-probability track), the level of the test stimulus starts clearly above the PSE, and it is adaptively varied after each response in the track so that the overall probability of "louder" responses approach a relatively high value (e.g., 71% "louder" responses). In the other track (i.e., the low-probability track), the initial level for the test stimulus starts clearly below the PSE, and the stimulus level was varied so that the "louder" responses approach a relatively low value (e.g., 29% "louder" responses). The final estimate of the PSE is the average of the stimulus levels at the 71 and 29% point of the psychometric function, estimated from the two tracks. This procedure adaptively adjusts the range of stimulus presentation and samples the test stimulus above and below the PSE with approximately equal probability. Traditionally, the high- and low-probability tracks are implemented using the staircase (transformed up-down) procedure (Levitt, 1971). This procedure will be referred to as the interleaved staircase (ISC) procedure in the following discussions.

As an alternative to the ISC procedure, one could also use interleaved Bayesian adaptive procedures for the estimation of the PSE. A number of Bayesian adaptive procedures have been proposed to estimate the psychometric function (e.g., Watson and Pelli, 1983; King-Smith and Rose, 1997; Kontsevich and Tyler, 1999). Since many of these procedures were developed to estimate performance thresholds, they contain algorithms to optimize the stimulus choices for threshold estimation, and the expected response probability (e.g., the probability of a "correct" response) is associated with how the performance threshold is defined. Shen and Richards (2012) proposed a hybrid procedure, namely, the updated maximum-likelihood (UML) procedure, between the Bayesian (or maximum-likelihood) procedure and the ISC procedure. This procedure allows the experimenter to explicitly control for the expected response probability, therefore the UML procedure can be implemented in an interleaved fashion with the two interleaved UML tracks corresponding to a high and a low response probability. We will refer to this modification of the UML procedure as the

[a)]Electronic mail: shen2@indiana.edu

interleaved maximum-likelihood (IML) procedure in the following discussion.

Both the ISC and IML procedures are designed to collect one PSE at a time by varying stimulus along only one dimension, i.e., the stimulus level. In many studies that involve loudness comparison, the PSE is commonly measured as a function of a second stimulus dimension, usually the frequency (e.g., Fletcher and Munson, 1933; Robinson and Dadson, 1956; Suzuki and Takeshima, 2004), bandwidth (e.g., Zwicker *et al.*, 1957; Verhey and Kollmeier, 2002), or duration (e.g., Zwislocki, 1969; Florentine *et al.*, 1996) of the test stimulus. The level required for a test stimulus to match the loudness of a standard stimulus as a function of the frequency of the test stimulus is commonly referred to as the equal-loudness contour (e.g., Fletcher and Munson, 1933), since any point along the contour represents the PSE. The current study investigates the possibility of estimating the PSE efficiently across multiple frequencies using a Bayesian adaptive procedure in an effort to improve the time efficiency of estimating the equal loudness-level contours. Instead of testing one test frequency at a time, this new procedure interleaves two adaptive tracks that estimate a high- and a low-percentage point on the psychometric function. Within each of the adaptive tracks, the test frequencies varied randomly from trial to trial. In the following discussion, we will refer to this Bayesian adaptive procedure as the interleaved equal-loudness contour (IELC) procedure. The availability of an efficient procedure would enable the estimation of equal-loudness contours from individual listeners, allowing individualized loudness control in a wide range of applications.

In the following, we will first describe the implementation of the IELC procedure and evaluate its test-retest reliability in Experiment 1 (Exp. 1). Then, results from Experiment 2 (Exp. 2) will be presented, in which the IELC procedure was compared to the ISC and IML procedures. It will be shown that the IELC procedure has demonstrated satisfactory efficiency and reliability. However, certain limitations of the IELC procedure have also been observed, and these limitations will be discussed.

II. THE INTERLEAVED EQUAL-LOUDNESS CONTOUR (IELC) PROCEDURE

A. Overview

The IELC procedure is designed to estimate the equal-loudness contour, i.e., the level of a narrowband test stimulus (e.g., a pure tone) needed to match the loudness of a standard stimulus. The procedure estimates a high-probability and a low-probability contour using two interleaved adaptive tracks. Each of the high- and low-probability contours is formulated as a set of candidate models, and each of the candidate models contains several model parameters to be estimated. The number of model parameters is kept relatively small (five in the current study) compared to the number of frequencies usually required to fit the equal-loudness contour if tested one frequency at a time. The IELC procedure, by modeling the equal-loudness contour using a small number of parameters, tries to take advantage of the covariance among the PSEs in adjacent frequency regions.

At the beginning of the IELC procedure, all candidate models have equal prior likelihoods and a prior distribution is assigned to the parameters of each candidate model. Following each trial, the posterior parameter distribution of each candidate model is updated via extended Kalman filtering (e.g., Shen *et al.*, 2014), and the likelihoods of the candidate models are updated according to Bayes' rule. The candidate model with the maximum-likelihood provides the estimate of the high- or low-probability contours and is used to select the stimulus for the next trial of the same track (see Fig. 1 for a summary of the computational steps between two consecutive trials).

B. Candidate models

In the current implementation of the IELC procedure, the frequency axis was represented by 11 discrete values logarithmically spaced between 250 and 4000 Hz. Each of the high- and low-probability contours was formulated as a set of candidate models, each of which was a piece-wise cubic interpolation curve ("pchip" function in MATLAB, see also Fritsch and Carlson, 1980) of five anchor points. The number of anchor points, $k_{\mathrm{IELC}} = 5$, was chosen according to pilot studies. The frequencies of the anchor points (i.e., anchor-point frequencies) were a subset of the 11 discrete frequencies. Therefore, there were a total of $M = \binom{5}{11} = 462$ unique sets of anchor points, corresponding to $M = 462$ candidate models. The prior likelihoods of the candidate models were set to be the same value (i.e., the prior likelihood for the mth candidate model is $l_{\mathrm{H},m,0} = 1/462$ for the high-probability track and $l_{\mathrm{L},m,0} = 1/462$ for the low-probability track). Each candidate model was determined by a set of five free parameters correspond to the levels of the five anchor points. The prior distribution of the model parameters was modeled as a five-dimensional Gaussian distribution. For the

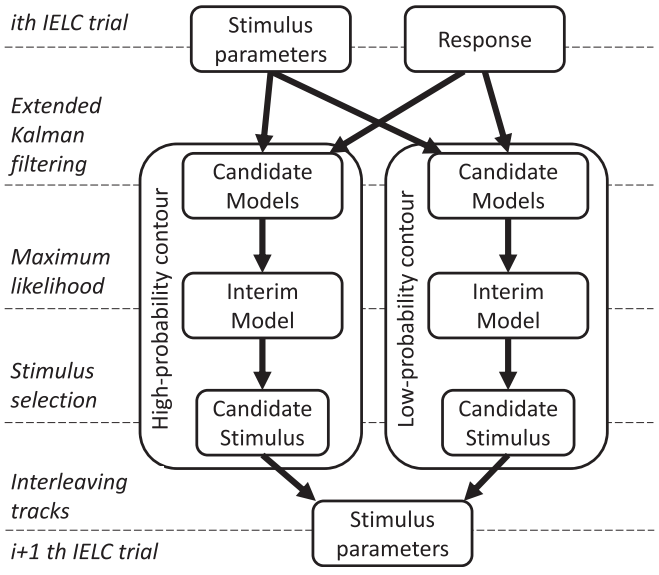

FIG. 1. Summary of the computational steps between two consecutive trials of an IELC run.

mth candidate model, the prior distribution had a mean of $\boldsymbol{\phi}_{\mathrm{H},m,0}$ (a 5 × 1 vector) and a covariance matrix of $\mathbf{P}_{\mathrm{H},m,0}$ (a 5 × 5 matrix) for the high-probability track, and a mean of $\boldsymbol{\phi}_{\mathrm{L},m,0}$ (a 5 × 1 vector) and a covariance matrix of $\mathbf{P}_{\mathrm{L},m,0}$ (a 5 × 5 matrix) for the low-probability track. The covariance matrices $\mathbf{P}_{\mathrm{H},m,0}$ and $\mathbf{P}_{\mathrm{L},m,0}$ were diagonal matrices, with the diagonal elements being the prior variances associated with the five anchor points.

The candidate models made predictions to the expected responses according to two psychometric functions, separately for the high- and low-probability tracks. For the high-probability track, it was assumed that there was a threshold associated with a 75% probability of "louder" responses. When the stimulus level was above the threshold, a response probability higher than 75% was expected, and when the stimulus level was below the threshold, a lower-than-75% "louder" response is expected. It was further assumed that the response probability was not expected to drop below 50% during the high-probability track. A logistic function was used to describe the psychometric function

$$p_{\mathrm{H},m}\left[f,L;\Theta_{\mathrm{H},m}(f;\boldsymbol{\phi}_{\mathrm{H},m,i}),\beta\right] = 0.5+0.5\left\{1+e^{-\beta\left[L-\Theta_{H,m}(f;\boldsymbol{\phi}_{\mathrm{H},m,i})\right]}\right\}^{-1}, \quad (1)$$

where $p_{\mathrm{H},m}$ is the probability of "louder" responses for the high-probability track predicted by the mth candidate model, f and L are the frequency and level of the test stimulus, $\Theta_{\mathrm{H},m}(f;\boldsymbol{\phi}_{\mathrm{H},m,i})$ is the predicted high-probability contour (i.e., the 75% threshold as a function of frequency f) from the mth candidate model (with model parameters $\boldsymbol{\phi}_{\mathrm{H},m,i}$) following the ith trial. The parameter β in Eq. (1) describes the slope of the psychometric function with a larger β value corresponding to a steeper psychometric-function slope. This parameter was set to $\beta=0.25$ in the current study.

For the low-probability track, the psychometric function was given by

$$p_{\mathrm{L},m}\left[f,L;\Theta_{\mathrm{L},m}(f;\boldsymbol{\phi}_{\mathrm{L},m,i}),\beta\right] = 0.5\left\{1+e^{-\beta\left[L-\Theta_{\mathrm{L},m}(f;\boldsymbol{\phi}_{\mathrm{L}m,i})\right]}\right\}^{-1}, \quad (2)$$

where $p_{\mathrm{L},m}$ is the probability of "louder" responses predicted by the mth candidate model, f and L are the frequency and level of the test stimulus, $\Theta_{\mathrm{L},m}(f;\boldsymbol{\phi}_{\mathrm{L},m,i})$ is the predicted low-probability contour (i.e., the 25% threshold as a function of frequency f) from the mth candidate model following the ith trial. The slope parameter $\beta=0.25$ was set to be the same as the slope parameter for the high-probability track [Eq. (1)].

The psychometric functions given in Eqs. (1) and (2) were not meant to model the perceptual or decisional processes underlying the loudness-comparison judgement, rather they were link functions between the stimulus level and response probability and were used to update the estimates of the high- and low-probability contours. The threshold parameter [$\Theta_{\mathrm{H},m}(f;\boldsymbol{\phi}_{\mathrm{H},m,i})$ in Eq. (1) or $\Theta_{\mathrm{L},m}(f;\boldsymbol{\phi}_{\mathrm{L},m,i})$ in Eq. (2)] and the slope parameter β defined the center and the width of the dynamic range of the psychometric function. When a stimulus was presented within the dynamic range, the threshold estimate was adjusted according to the collected response (the parameter-updating procedure will be described in detail later). On the other hand, when a stimulus was presented outside of the dynamic range (too close to 100% or below 50% of "louder" responses for the high-probability track or too close to 0% or above 50% of "louder" responses for the low-probability track), the threshold estimate would not be affected by the response. For shallower assumed psychometric function slopes (i.e., smaller values of β), the threshold would be updated with larger steps, while for steeper slopes (i.e., larger values of β), the threshold would be updated with smaller steps.

C. Model update

Following the ith trial, the response collected r_i ($r_i=1$ for a "louder" response, otherwise $r_i=0$) was used to update the parameters of all M candidate models for both the high- and low-probability contours via extended Kalman filter. Kalman filtering is a technique to provide treatment to noisy observations from a dynamical system. Assuming the system can be modeled as a linear system for a state variable, the noisy observation can be considered as a function of the state variable (i.e., the observation function) plus an additive observational noise. After each observation, the state variable and its covariance matrix can be updated according to the deviation between the measured observation and the expected observation predicted by the model. In the current application, for the mth model, the state variable was $\boldsymbol{\phi}_{\mathrm{H/L},m,i}$, the observation was r_i, and the observation function was the psychometric function [Eqs. (1) and (2)]. Since the observation was in the form of binary responses (i.e., "0" or "1"), the observation noise was assumed to follow the Bernoulli distribution.[1] Because the psychometric function was not a linear function, extended Kalman filtering was used, which involved local linearization of the psychometric function.

For the mth candidate model, the mean $\boldsymbol{\phi}_{\mathrm{H/L},m,i}$ and covariance matrix $\mathbf{P}_{\mathrm{H/L},m,i}$ of the posterior parameter distribution following the ith trial were given by

$$\boldsymbol{\phi}_{\mathrm{H/L},m,i}=\boldsymbol{\phi}_{H/L,m,i-1}+\mathbf{K}(r_i-\mu_m), \quad (3)$$

and

$$\mathbf{P}_{\mathrm{H/L},m,i}=\mathbf{P}_{\mathrm{H/L},m,i-1}-\mathbf{K}\cdot\mathbf{J}\cdot\mathbf{P}_{\mathrm{H/L},m,i-1}, \quad (4)$$

where

$$\mathbf{K}=\mathbf{P}_{\mathrm{H/L},m,i-1}\cdot\mathbf{J}^{\mathrm{T}}\cdot\left[\mathbf{J}\cdot\mathbf{P}_{\mathrm{H/L},m,i-1}\cdot\mathbf{J}^{\mathrm{T}}+\sigma_m^2\right]^{-1}. \quad (5)$$

Equation (3) is the discrete-time state model of the variable $\boldsymbol{\phi}_{\mathrm{H/L},m,i}$, and it describes the state transition from the i - 1th trial to the ith trial. The term $(r_i-\mu_m)$ in Eq. (3) represents the deviation from the observed response to the expected response predicted from the mth candidate model. The predicted mean response probability μ_m was calculated as

$$\mu_m = \begin{cases} p_{\mathrm{H/L},m}\left[f_i, L_i; \Theta_{\mathrm{H/L},m}(f_i; \phi_{\mathrm{H/L},m,i-1}), \beta\right], & \text{for } r_i = 1 \\ 1 - p_{\mathrm{H/L},m}\left[f_i, L_i; \Theta_{\mathrm{H/L},m}(f_i; \phi_{\mathrm{H/L},m,i-1}), \beta\right], & \text{for } r_i = 0', \end{cases} \quad (6)$$

where f_i and L_i are the frequency and level of the test stimulus presented on the ith trial, respectively, and $p_{\mathrm{H/L},m}[f_i, L_i; \Theta_{\mathrm{H/L},m}(f_i; \phi_{\mathrm{H/L},m,i-1}), \beta]$ represents the psychometric function given in Eq. (1) for the high-probability track or Eq. (2) for the low-probability track. The expected variance of the observational noise is predicted by the mth candidate model σ_m^2 followed the Bernoulli distribution. That is

$$\sigma_m^2 = \mu_m(1 - \mu_m). \quad (7)$$

The matrix $\mathbf{J}$ in Eqs. (4) and (5) is a 1×5 matrix that represents the Jacobian of the psychometric function

$$\mathbf{J} = \left[\frac{\partial p_{\mathrm{H/L},m}}{\partial \phi_{\mathrm{H/L},m,i-1,1}}, \frac{\partial p_{\mathrm{H/L},m}}{\partial \phi_{\mathrm{H/L},m,i-1,2}}, \ldots, \frac{\partial p_{\mathrm{H/L},m}}{\phi_{\mathrm{H/L},m,i-1,5}}\right], \quad (8)$$

where $\phi_{\mathrm{H/L},m,i-1,n}$ is the nth parameter out of all five parameters in $\boldsymbol{\phi}_{\mathrm{H/L},m,i-1}$. The Jacobians were calculated as a part of the local linearization process in the extended Kalman filtering algorithm.

Once $\boldsymbol{\phi}_{\mathrm{H/L},m,i}$ and $\mathbf{P}_{\mathrm{H/L},m,i}$ were updated, the likelihood for the mth candidate model was calculated as

$$\ell_{\mathrm{H/L},m,i} = \frac{\ell_{\mathrm{H/L},m,i-1} \cdot \mu_m}{\sum_m \ell_{\mathrm{H/L},m,i-1} \cdot \mu_m}. \quad (9)$$

The candidate model with the maximum likelihood provided the interim estimates of the threshold contours (referred to as the "Interim Model," see Fig. 1). Using the subscript "max" to denote the candidate model with the maximum likelihood, the interim estimate of the high- or low-probability contours can be written as $\Theta_{\mathrm{H/L,max}}(f; \boldsymbol{\phi}_{\mathrm{H/L,max},i})$.

In a typical run of the IELC procedure, the likelihood of a few candidate models stood out from the rest of the candidate models after the first a few trials. As additional data were collected, the likelihood of most candidate models dropped rapidly, while the likelihood of a small number of candidate models (typically two or three) remained high and comparable to one another. The candidate model with the maximum likelihood typically switched this small set of candidate models. Which set of candidate models is associated with high likelihood depends on the shape of underlying equal-loudness contour and is likely to vary from individual to individual.

D. Stimulus selection

Following the ith trial, the frequency and level of the test stimulus on the $i+1$th trial was selected using the following procedure. First, the stimulus frequency f_{i+1} was randomly selected from the 11 discrete frequencies with equal probability.[2] Then, the stimulus level L_{i+1} was selected at $\Theta_{\mathrm{H,max}}(f_{i+1}; \boldsymbol{\phi}_{\mathrm{H},m,i})$ if the $i+1$th trial belonged to the high-probability track and at $\Theta_{\mathrm{L,max}}(f_{i+1}; \boldsymbol{\phi}_{\mathrm{L},m,i})$ if the $i+1$th trial belonged to the low-probability track. The high- and low-probability tracks were randomly interleaved, therefore, L_{i+1} was sampled at $\Theta_{\mathrm{H,max}}(f_{i+1}; \boldsymbol{\phi}_{\mathrm{H},m,i})$ and $\Theta_{\mathrm{L,max}}(f_{i+1}; \boldsymbol{\phi}_{\mathrm{L},m,i})$ with equal probability. Because $\Theta_{\mathrm{H,max}}(f_{i+1}; \boldsymbol{\phi}_{\mathrm{H},m,i})$ and $\Theta_{\mathrm{L,max}}(f_{i+1}; \boldsymbol{\phi}_{\mathrm{L},m,i})$ correspond to the expected stimulus levels associated with the 75 and 25% response probabilities, respectively, sampling stimuli at these levels ensured an overall response probability of 75% for the high-probability track and 25% for the low-probability track.

E. *Post hoc* model fit

For each IELC run, following data collection, the stimulus parameters, i.e., the frequencies and levels of the test stimuli, and the responses from all trials, from both the high- and low-probability tracks, were grouped under the 11 test frequencies. For each test frequency, a logistic regression between the test levels L_i (as the independent variable) and responses r_i (as the dependent variable) was conducted. Specifically, the probability of the "louder" responses ($r_i = 1$) was modeled as

$$p(r_i = 1) = \left[1 + e^{-b(L_i - \mathrm{PSE})}\right]^{-1}, \quad (10)$$

where the PSE and the parameter b are the two parameters of the logistic model. The PSE was estimated by fitting the logistic model to the experimental data. The above process was repeated for each of the test frequencies, leading to 11 PSE estimates.

Besides the PSEs, the variability of the PSE estimates was also derived for each IELC run using a bootstrap procedure. The bootstrap procedure involved repeatedly resampling the data and refitting the logistic model as in Eq. (10). For each repetition, 500 trials were drawn from the data with replacement. Based on the redrawn data, the PSEs at the 11 test frequencies were estimated. This was repeated for 1000 times, and the 68% confidence interval for the PSE was derived at each test frequency.

III. EXPERIMENT 1: THE TEST-RETEST RELIABILITY OF THE IELC PROCEDURE

A. Methods

1. Subjects

Six listeners (S1–S6, five females) between 20 and 29 years of age were recruited for the current experiment. All listeners had hearing thresholds that were 15 dB hearing level (HL) or better between 250 and 4000 Hz, i.e., the range of frequencies within which the equal-loudness contours were estimated. For each listener, the average hearing thresholds at 0.5, 1, and 2 kHz were calculated, and the ear with better (i.e., lower) average hearing threshold was tested in the experiment. All listeners were naive to loudness comparison experiments. The experimental protocol was approved by the Institutional Review Board at Indiana University Bloomington, and informed consent was obtained from all listeners before data collection.

2. Stimuli

In the current experiment, two pure tones were presented sequentially to the listener on each trial, separated by an inter-stimulus interval of 500 ms. The duration of the pure tones was 300 ms, including 20 ms raised-cosine onset/offset ramps. One of the pure tones, determined at random, was the standard stimulus, which was a 1 kHz pure tone at 60 dB sound pressure level (SPL). The other pure tone was the test stimulus, and its frequency and level were determined before each trial by the IELC procedure. The listener was instructed to select the tone that sounded louder. The response from the ith trial $r_i = 1$ if the test stimulus was selected, while $r_i = 0$ if the standard stimulus was selected.

All stimuli were presented at a sampling rate of 22 050 Hz. They were presented to the listeners via a 24-bit soundcard (Microbook II, Mark of the Unicorn, Inc.) using headphones (HD280 Pro, Sennheiser electronic GmbH and Co. KG). During the experiment, the listeners were seated in a sound-attenuating booth.

3. Procedure

The current experiment consisted of two IELC runs, i.e., a test and a retest. Each of the IELC runs included 200 trials. The prior parameter distributions for the candidate models of the first IELC run were configured so that the prior means were 60 dB SPL (for all elements in $\boldsymbol{\phi}_{\mathrm{H/L},m,0}$ with $m = 1$ to M) and the prior variances were 1600 (for all diagonal elements in $\mathbf{P}_{\mathrm{H/L},m,0}$ with $m = 1$ to M). For the second IELC run, the prior parameter distributions were identical compared to the first IELC run, except that any anchor points in any of the candidate models with frequencies less than or equal to 500 Hz were assigned with a prior mean of 70 dB SPL. The difference in the prior parameter distributions of the two IELC runs was introduced to investigate the influence of prior configurations. The experiment was conducted in a single test session, no prior training was provided, and the listener was not given any feedback following each response. Naive listeners with no training were tested in the current experiment to test the IELC procedure under a challenging scenario. Moreover, the aim of the IELC procedure was to improve the time efficiency of data collection. If substantial training is required prior to an IELC run in order to achieve satisfactory reliability, then the actual time-saving may be limited. Therefore, it is worth verifying the test-retest reliability of the IELC procedure using naive listeners.

B. Results

As described earlier, each IELC run contained a high- and a low-probability track corresponding to two iso-performance contours at the 75 and 25% response probability, respectively. Figure 2 plots the estimated high- and low-probability contours (as upward and downward triangles, respectively) for the two IELC runs and two representative listeners (S1 and S5). The estimated high-probability contours were higher in level compared to the low-probability contour. The distance between the high- and low-probability contours was as large as 20 dB. Most stimuli sampled by the IELC procedure were in the adjacent regions near the two contours. Moreover, as expected from the responses probabilities associated with the two contours, stimuli presented above the high-probability contour mostly resulted in "louder" responses (circles) while stimuli presented below the low-probability contour mostly resulted in "weaker" responses (crosses). This means that the estimated contours were able to capture the dynamic range of the psychometric function at each of the test frequencies.

For the examples illustrated in Fig. 2, the estimated high- and low-probability contours were consistent across the two IELC runs. The stimuli sampled during the two IELC runs, though not identical, occupied similar regions of the stimulus space. This indicates that the IELC runs were able to converge to similar estimates of the high- and low-probability contours following 200 trials of testing.

Figure 3 plots the estimated equal-loudness contours from the first and second IELC runs for all six listeners (one panel for each listener). For most listeners and most frequencies, the PSE estimates converged well, with the 68% confidence intervals (error bars) typically within 10 dB. Occasionally, certain test frequencies were sampled only very few times. This led to relatively large confidence intervals (see the first IELC run for S5 as an example). Large individual differences in the shape of the equal-loudness contour were observed. On the other hand, the estimated equal-loudness contours from the two IELC runs resembled each other well. The root-mean-square (rms) deviation between the two estimated equal-loudness contours ranged from 3.0 to 5.5 dB among the six listeners (with a mean of 4.3 dB and a standard deviation of 0.8 dB).

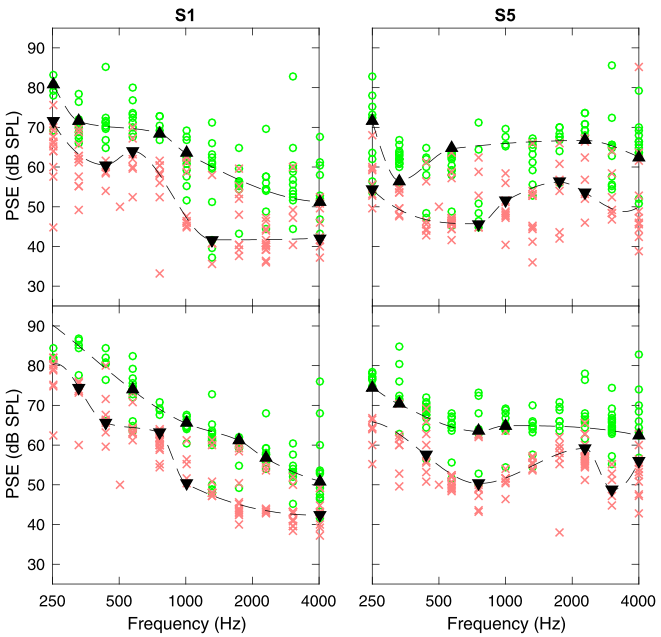

FIG. 2. (Color online) The stimulus distributions and estimated high- and low-probability contours for the two IELC runs (in rows) and two representative listeners (in columns) in Exp. 1. In each panel, circles indicate the test stimuli resulted in "louder" responses while crosses indicate the test stimuli resulted in "weaker" responses. The estimated high- and low-probability contours at the end of the 200-trial IELC run are plotted using dashed curves with the anchor points indicated using upward and downward triangles, respectively.

To investigate whether the test-retest reliability was affected by the difference in the prior parameter distributions, the rms deviations between the two IELC runs were calculated separately across frequencies where the 10-dB difference in the prior mean was implemented (i.e., three frequencies below 500 Hz) and across frequencies where the identical prior mean was used (i.e., eight frequencies above 500 Hz) for each listener. A paired t-test indicated no significant effect of whether the prior distributions agreed across runs ($t_5 = -1.39$, $p = 0.224$). This suggests that the excellent test-retest reliability of the IELC procedure cannot be explained by the similarities in prior configurations.

IV. EXPERIMENT 2: COMPARISONS AMONG THREE INTERLEAVED ADAPTIVE PROCEDURES

A. General methods

Three different psychophysical procedures to estimate the equal-loudness contours were compared in Exp. 2. Following Exp. 1, all six listeners from Exp. 1 participated in Exp. 2. The listeners' task and stimuli were identical to those in Exp. 1. Three of the six listeners started with two runs of the ISC procedure, followed by two runs of the IML procedure and a third run of the IELC procedure (counting the two IELC runs in Exp. 1 as the first and second). For the other three listeners, the experiment began with two runs of the IML procedure, followed by two runs of the ISC procedure and a third run of the IELC procedure. The third IELC run was tested at least one week after the first two IELC runs. It was included to check for the test-retest reliability of the IELC procedure across a longer time interval.

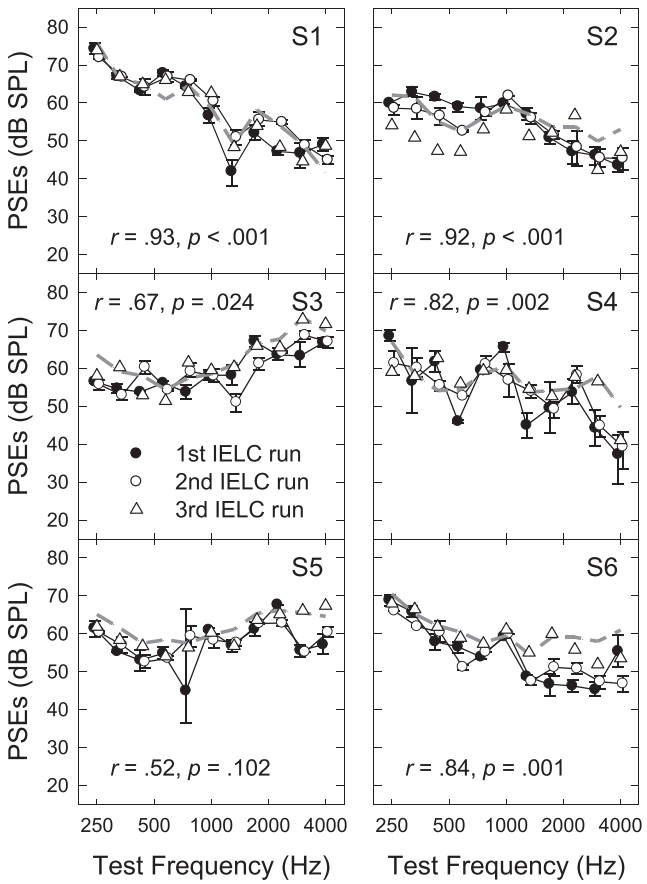

FIG. 3. The equal-loudness contours estimated from the three IELC runs. Results for each individual listener (S1–S6) are arranged in respective panels. Within each panel, the estimates from the three IELC runs are indicated using different symbols, and the dashed line indicates the best estimate of the equal-loudness contours by pooling the data from Exps. 1 and 2 together. Error bars indicate the 68% confidence intervals. The symbols were horizontally shifted and the error bars are not shown for the third IELC run for visual clarity. The inserted text in each panel indicates the correlation coefficient between the first and second IELC runs.

B. The ISC procedure

For each run of the ISC procedure, all test frequencies except for 1 kHz were tested in random order. The PSE for each of the test frequencies was estimated using one experimental block. Each experimental block consisted of two randomly interleaved adaptive tracks, i.e., the high- and low-probability tracks. For the high-probability track, the initial stimulus level was 80 dB SPL, and a 2-down, 1-up rule was used to manipulate the stimulus level, corresponding to the 71% point on the psychometric function. For the low-probability track, the initial stimulus level was 40 dB SPL, and a 1-down, 2-up rule was used, corresponding to the 29% point on the psychometric function. For both tracks, the initial step size for varying the stimulus level was 10 dB, which was reduced to 5 dB after the first two reversals of the adaptive track. The final step size of 5 dB was chosen based on pilot testing. The pilot results indicated fairly shallow psychometric functions, therefore a step size of 2 dB typically used for pure-tone detection experiments would be too small for the current loudness-comparison task. The adaptive tracks terminated when the number of reversals in both of the tracks became eight or larger. Each experimental block typically consisted of approximately 50 trials.

C. The IML procedure

For each run of the IML procedure, all test frequencies except for 1 kHz were tested in random order. The PSE for each of the test frequencies was estimated using one experimental block. Each experimental block consisted of two randomly interleaved adaptive tracks, and each adaptive track followed the UML procedure (Shen and Richards, 2012; Hatzfeld *et al.*, 2015; Doire *et al.*, 2017) and was implemented using the UML toolbox for the MATLAB platform (Shen *et al.*, 2015).

The UML procedure formulates the psychometric function using a few model parameters. For each of these parameters, there are one or two optimal stimulus strengths that would minimize the expected variance for the parameter. These stimulus strengths are called sweet-points. At the beginning of the UML procedure, a prior distribution is defined for the psychometric-function parameters. Following each trial, the posterior parameter distribution is updated according to Bayes' rule. The mean of the posterior distribution provides an interim estimate of the psychometric function. Then, the sweet-points are re-estimated based on the interim psychometric function, and the stimulus for the next trial is selected as one of the sweet-points using an adaptive procedure.

For a test frequency, each of the two interleaved UML tracks was implemented with a logistic psychometric function

$$p(r_i = 1) = \gamma + (1 - \gamma - \lambda)\left[1 + e^{-b(L_i - \mathrm{PSE})}\right]^{-1}, \quad (11)$$

where b is the slope parameter of the logistic psychometric function with a larger b value indicating a steeper slope; γ is the distance between the lower plateau of the psychometric function and 0 (i.e., the response probability does not reach 0% even at a very low level); λ is the distance between the upper plateau of the psychometric function and 1 (i.e., the response probability does not reach 100% even at a very high level). The prior distribution for PSE was a Gaussian distribution with a mean of 60 dB SPL and a standard deviation of 10 dB; the prior distribution for $\log_{10}(b)$ was a Gaussian distribution with a mean of -1 and a standard deviation of 0.5; the prior distributions for γ and λ were two uniform distributions spanning 0.02 and 0.2. The UML procedure updated the posterior distributions of these four parameters following each trial according to Bayes' rule, and the mean of the posterior parameter distributions formed the interim psychometric-function estimate.

For the psychometric function defined in Eq. (11), there was one sweet-point associated with the PSE and two sweet-points associated with b, one above and one below the PSE sweet-point (Shen and Richards, 2012). The levels associated with these three sweet-points were iteratively computed based on the interim psychometric-function estimate following every trial. In addition, two additional sweet-points were derived, one below the lower b sweet-point and one above the upper b sweet-point, so that the five sweet-points were evenly spaced in level. The lowest and highest sweet-points were implemented to provide estimates of γ and λ, respectively. This implementation deviated from the original UML procedure described by Shen and Richards (2012) and Shen *et al.* (2015), where the sweet-points for γ and λ were positioned at the lower and upper limits of the stimulus space. This previous approach could lead to large level jumps when the stimulus level moves between the b sweet-points and their adjacent γ/λ sweet-point. This type of abrupt level jumps was prevented by the evenly spaced sweet-points.

For the high-probability track, the initial stimulus level was 80 dB SPL, and a 2-down, 1-up sweet-point selection rule was used. That is, the stimulus level was shifted to one sweet-point lower following two consecutive "louder" ($r = 1$) responses, unless the stimulus level was already at the lowest sweet-point. On the other hand, the stimulus level was shifted to one sweet-point higher following a single "weaker" ($r = 0$) response, unless the stimulus level was already at the highest sweet-point. This sweet-point selection rule corresponded to an expected response probability of 71%. For the low-probability track, the initial stimulus level was 40 dB SPL, and a 1-down, 2-up sweet-point selection rule was used, corresponding to an expected response probability of 29%. The adaptive tracks terminated when the number of reversals in both of the tracks became eight or larger. Each experimental block typically consisted of approximately 50 trials.

D. *Post hoc* model fit

For each equal-loudness contour estimate in Exp. 2 (via the ISC, IML, or IELC procedure), the data were grouped under all test frequencies (10 test frequencies for the ISC and IML procedures and 11 test frequencies for the IELC procedure). For each test frequency, the PSE was estimated through logistic regression similar to Exp. 1. Although the ISC and IML procedures were able to produce estimates of the PSE directly, the *post hoc* logistic regression was used to ensure that any differences among the procedures were not caused by differences in how the PSEs were estimated after the data were collected. Similar to Exp. 1, the variability of the PSE estimates was derived for each test run using a bootstrap procedure. The bootstrap procedure involved 1000 repetitions. For each repetition, 500 trials were drawn from the collected data with replacement. The redrawn data were grouped by test frequencies, and for each test frequency the logistic model [i.e., Eq. (10)] was fitted to the data. The 68% confidence interval for the PSE at each test frequency was derived based on the PSE estimates from the 1000 bootstrap repetitions.

Besides the test-retest reliability, the accuracies of the three adaptive procedures were also evaluated. Since the "true" equal-loudness contours were not available, the absolute accuracies could not be calculated. Instead, for each listener, the trial-by-trial data across the two experiments and all three adaptive procedures were pooled together and the pooled data was used to derive a "best estimate" of the equal-loudness contour using the same logistic regression procedure described above. The accuracies of the adaptive procedures were assessed using the rms deviation from individual estimate of the equal-loudness contour to the best estimate. Since the total number of trials for the IELC procedure (600 trials) was less than the ISC and IML procedures (approximately 1000 trials for each procedure), if the best estimate is based on the pooled data directly, the resulting best estimate may be bias toward the procedures with more trials in the pool. To enable fair comparisons, the best estimate of the equal-loudness contour was derived via a bootstrap approach. One thousand bootstrap repetitions were conducted. Within each repetition and for each of the IELC, ISC, and IML procedures, the pooled data consisted of 1000 trials randomly drawn from the experimental data with replacement. One equal-loudness contour was then estimated based on the pooled data via logistic regression. The reported best estimate of the equal-loudness contour was the average of the derived equal-loudness contours across the 1000 bootstrap repetitions.

E. Results

Figure 4 plots the estimated equal-loudness contours from the first and second ISC runs. The rms deviation between the two estimated equal-loudness contours ranged from 1.8 to 8.1 dB among the six listeners (with a mean of 4.5 dB and a standard deviation of 2.4 dB). Figure 5 plots the estimated equal-loudness contours from the first and second IML runs. The rms deviation between the two estimated equal-loudness contours ranged from 2.9 to 9.6 dB among

the six listeners (with a mean of 5.2 dB and a standard deviation of 2.8 dB). These rms deviations between the test and retest for the ISC and IML procedures were similar to that found in Exp. 1 for the IELC procedure.

The effect of test procedures (i.e., the IELC, ICS, and IML procedures) on the rms deviations between the test and retest runs were investigated using a repeated measures analysis of variance (ANOVA). The first and second IELC runs from Exp. 1 were used in this analysis. No significant effect of test procedures on the test-retest deviation was found [$F(2, 10) = 0.47$, $p = 0.639$]. This suggests that the test-retest reliability of the three procedures were comparable even when more trials were included in the ISC (500 trials per run) and the IML (500 trials per run) procedures than the IELC procedure (200 trials per run).

To investigate the test-retest reliability of the IELC procedure across a relatively long time interval, the equal-loudness contours estimated from the first and third IELC runs were compared. These two IELC runs were tested at least one week apart. The rms deviation between the first and third IELC runs ranged from 2.9 to 8.0 dB among the six listeners (with a mean of 5.7 dB and a standard deviation of 1.8 dB). The rms deviation between the first and third IELC runs was not significantly larger than the rms deviation between the first and second IELC runs ($t_5 = -1.56$, $p = 0.180$). This suggests that the general shape of the equal-loudness contour was relatively stable for each listener, and the variability in the equal-loudness contours across individual listeners was not dominated by the measurement error.

The accuracies of the IELC, ISC, and IML procedures were evaluated against the best estimate of the equal-loudness contours. For each of the six listeners, the rms error was calculated as the rms deviation from the best estimate for the listener. Figure 6 summarizes the average rms errors across listeners for all seven estimates of the equal-loudness contours (i.e., three IELC runs, two ISC runs, and two IML runs). The average rms errors were typically between 3 and 4 dB for all IELC, ISC, and IML runs in Exp. 2, but the two IELC runs in Exp. 1 exhibited higher average rms errors. A repeated measure ANOVA treating the various experimental runs as the independent variable and the rms error as the dependent variable revealed a significant effect of experimental runs [$F(6, 30) = 3.71$, $p = 0.007$]. *Post hoc* pair-wise comparisons suggests that the rms error for the first IELC run was marginally higher than that for the first IML run ($p = 0.077$, Bonferroni corrected). No other pairs of experiment runs led to significant differences. When repeating the ANOVA analysis with the data from Exp. 1 removed, the effect of experimental runs on the rms error was no longer significant [$F(4, 20) = 1.40$, $p = 0.271$]. It is possible that the higher rms errors for the first two IELC runs were related to the fact that all listeners started these two IELC runs without training. Once the listeners became familiarized with the experimental task, the rms errors for the third IELC run (tested at the end of Exp. 2) were similar to those of the ISC and IML procedures.

Since all three procedures were designed to adaptively sample stimuli associated with either a high or a low response probability, the tested stimuli from each of the

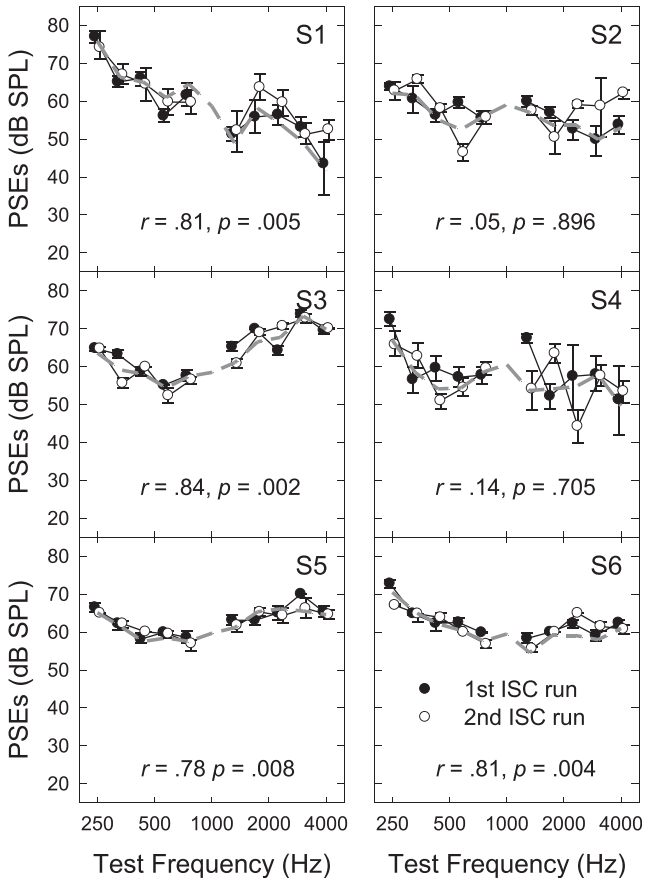

FIG. 4. The equal-loudness contours estimated from the two ISC runs. Results for each individual listener (S1–S6) are arranged in respective panels. Within each panel, the estimates from the two ISC runs are indicated using different symbols, and the dashed line indicates the best estimate of the equal-loudness contours by pooling the data from Exps. 1 and 2 together. Error bars indicate the 68% confidence intervals. The symbols were horizontally shifted for visual clarity. The inserted text in each panel indicates the correlation coefficient between the two ISC runs.

procedures should not be scattered evenly across the stimulus space. Rather, they should concentrate in areas that follow the shape of the equal-loudness contour across test frequencies but maintain a certain distance (in level) from the equal-loudness contour. To compare the stimulus-sampling patterns of the IELC, ISC, and IML procedures, the level differences from the trial-by-trial stimuli to the best estimate of the equal-loudness contours were calculated for each of the procedure (600 trials for the IELC procedure, and approximately 1000 trials for each of the ISC and IML procedures). Figure 7 summarizes the distributions of the sampled stimuli over the deviations from the best estimate. For two of the listeners (S1 and S2), the three procedures sampled stimuli in a similar fashion, with the most trials being concentrated near the PSE (i.e., near a deviation of 0 dB). For the rest of the listeners, the stimuli sampled by the IELC procedure were bimodally distributed. Regions just above or just below the PSE were sampled more frequently than the PSE itself. In contrast, the ISC and IML procedures sampled stimuli most frequently near the PSE. For listeners S3, S5, and S6, the IML procedure sampled within a very narrow region surrounding the PSE. Therefore, among the three procedures, the IELC procedure was the most

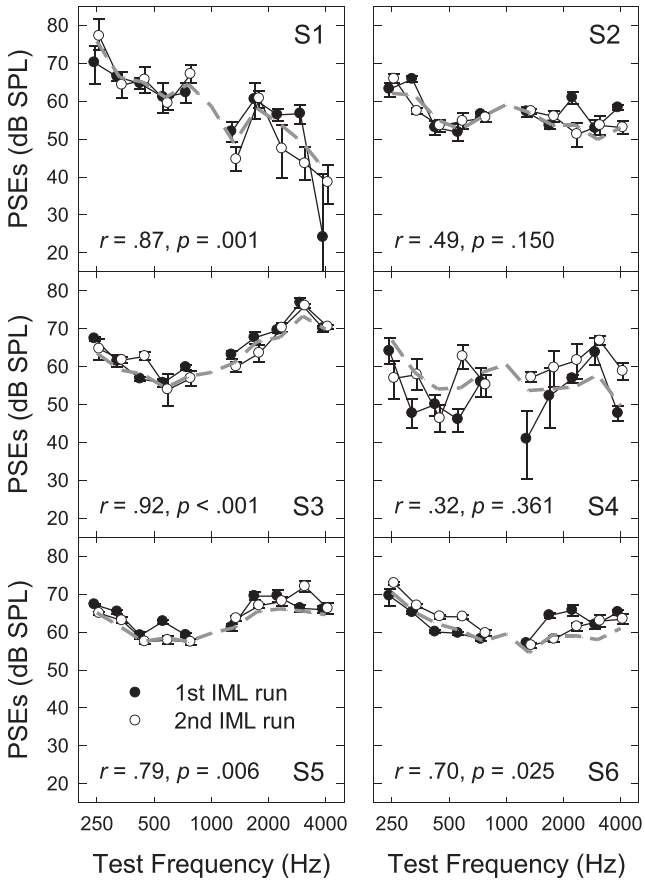

FIG. 5. The equal-loudness contours estimated from the two IML runs. Results for each individual listener (S1–S6) are arranged in respective panels. Within each panel, the estimates from the two IML runs are indicated using different symbols, and the dashed line indicates the best estimate of the equal-loudness contours by pooling the data from Exps. 1 and 2 together. Error bars indicate the 68% confidence intervals. The symbols were horizontally shifted for visual clarity. The inserted text in each panel indicates the correlation coefficient between the two IML runs.

successful in preventing the situations where the listeners were forced to compare the standard and test stimuli of very similar loudness.

The differences in the stimulus distributions reflected the different stimulus selection algorithms within the three procedures. For the IELC procedure, the contours associated with 75 and 25% responses (i.e., the high- and low-probability contours) are iteratively updated and the stimuli are sampled along the interim estimate of the high- and low-probability contours. On the other hand, although the interleaved tracks in the ISC and IML procedures converge towards average response probabilities of 71 and 29%, the stimuli are not sampled directly at the test levels associated with these response probabilities. For the ISC procedure, the test level steps up and down in the region of the targeted probability for each adaptive track. As a result, the stimuli sampled by the high- and low-probability tracks exhibit a fair amount of overlap, causing the region between the two targeted response probabilities, i.e., near the PSE, to be sampled most frequently. For the IML procedure, the test level is always on one of the sweet-points. These sweet-points are not the levels associated with the targeted response probabilities; rather they are the optimal points to estimate the

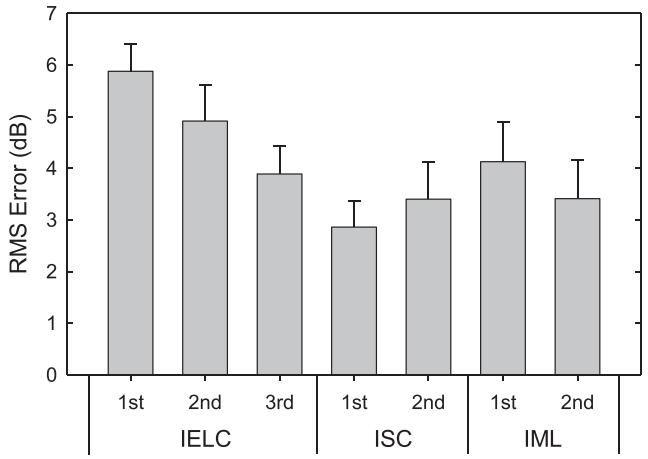

FIG. 6. The average rms errors across the six listeners for individual estimates of the equal-loudness contours. The IELC runs are shown in the first three bars from the left, the middle two bars are for the ISC runs, and the other two bars on the right are for IML runs. Error bars indicate one standard error of the mean.

parameters of the psychometric function. During each adaptive track of the IML procedure, the test level moves up and down the sweet-points to maintain an overall response probability close to the targeted probability. The two interleaved tracks share the same set of sweet-points. Because the sweet-point associated with the PSE are sampled frequently by both of the adaptive tracks, most test stimuli were distributed near the PSE.

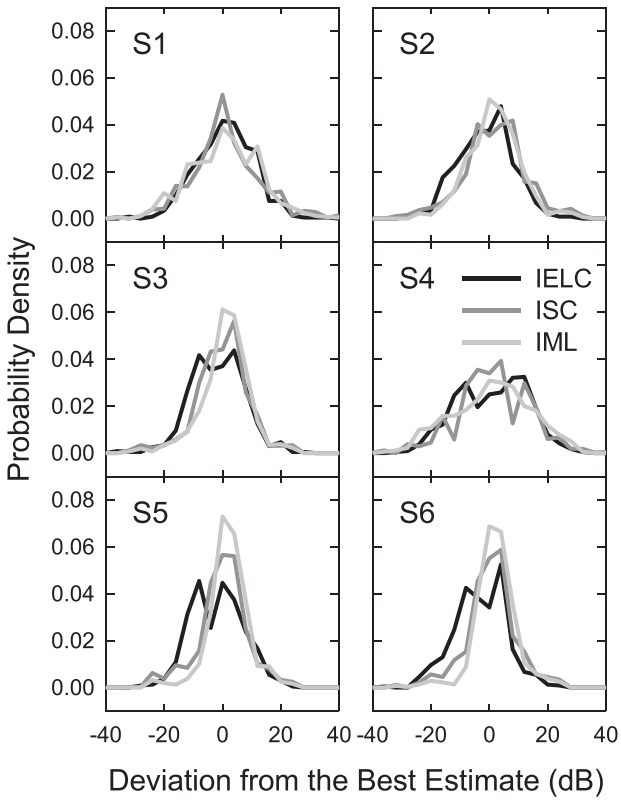

FIG. 7. The distributions of the stimuli sampled by the IELC, ISC, and IML procedures across the deviation (in level) from the best estimates of the equal-loudness contours for individual listeners (in separate panels).

V. GENERAL DISCUSSION

A. Performance of the IELC procedure

The current study extends upon an interleaved adaptive technique for measuring the PSEs (i.e., the ISC procedure, Jesteadt, 1980) in an effort to improve the time efficiency and reliability in estimating the equal-loudness contours. These extensions included two main approaches. For the first of these approaches, an interleaved Bayesian adaptive procedure, i.e., the IELC procedure, has been developed, which allows the simultaneous estimation of the equal-loudness contour across frequencies instead of measuring the PSEs one frequency at a time. This procedure gains its efficiency by assuming that the PSEs in the adjacent frequency regions are correlated, so that the equal-loudness contour could be represented using a relatively small number of model parameters. By this formulation, a response collected at a certain test frequency also provides information on the shape of the equal-loudness contour in nearby frequencies. It has been demonstrated in the current study that the IELC procedure can achieve excellent test-retest reliability using only 200 trials (approximately 20 min of testing time).

Although the IELC procedure has demonstrated satisfactory reliability and efficiency, it also has a number of major limitations because the assumptions underlying the procedure may not always be valid. One assumption is that the equal-loudness contour to be estimated is relatively smooth and it can be modeled as the piece-wise cubic interpolation curve across five anchor points. It is known that the hearing threshold exhibits fine structures when measured using pure-tone stimuli (i.e., fluctuations in threshold when measured in small frequency steps), reflecting local fine mechanical perturbations along cochlear partitions (e.g., Epp *et al.*, 2010). These fine structures can also be observed in low-intensity loudness measurements (Mauermann *et al.*, 2004). Therefore, if the equal-loudness contours possess similar fine fluctuations over frequency, the modeled contours using five anchor points would not be sufficient to capture its local details.

Even under conditions where the influences from the cochlear fine structure may be small (e.g., at high intensities), the equal-loudness contour may still exhibit a level of complexity that cannot be captured by the modeled threshold contour, or it may contain local details that do not align with the discrete frequency axis implemented in the IELC procedure. In order to demonstrate this potential limitation, a series of Monte-Carlo simulations were conducted. In these simulations, the IELC procedure was conducted on simulated listeners, who responded to the stimuli according to their "true" equal-loudness contours.

The "true" equal-loudness contours were generated using the following procedure. First, k_{true} frequencies were determined, including 250 Hz, 4000 Hz, and k_{true}–2 other frequencies randomly drawn from a uniform distribution spanning 250 and 4000 Hz (open interval). Note that the randomly drawn frequencies were not required to align with any of the 11 discrete frequencies used in the IELC procedure. Second, k_{true} levels were randomly drawn from a uniform distribution spanning 40 and 80 dB SPL, corresponding to the k_{true} frequencies. This created k_{true} "true" anchor points. Finally, the "true" equal-loudness contour was calculated as the piece-wise cubic interpolation curve across the k_{true} "true" anchor points. According to this construction procedure, the "true" equal-loudness contours typically varied wildly across frequency, much more than what one would encounter in human listeners. The complexity, i.e., the amount of local details, of the "true" equal-loudness contour was controlled by the value of k_{true}. As k_{true} increased, the "true" equal-loudness contour became more complex. To assess the effect of the model complexity, the number of anchor points in the IELC procedure, k_{IELC}, were also varied systematically.

For the current simulations, the values of k_{true} were 3, 4, 5, 6, and 7, and the values of k_{IELC} were from 4 to 10. For each combination of k_{true} and k_{IELC}, 100 simulated IELC runs were conducted. For each simulated IELC run, a "true" equal-loudness contour was independently constructed. On a given simulated trial, the PSE at the test frequency was evaluated along the "true" equal-loudness contour, and the response was simulated according to the psychometric function as specified in Eq. (10). A relative shallow slope was assumed for the simulated psychometric function ($b = 0.25$). For each simulated IELC run, the deviations between the estimated and "true" equal-loudness contours were evaluated at 201 logarithmically spaced frequencies from 250 to 4000 Hz and summarized into a rms error.

Figure 8 plots the average rms error as functions of k_{IELC} for the three values of k_{true}. As the complexity of the "true" equal-loudness contour increases, the rms error increases. That is, for a given number of anchor points in the IELC procedure, the accuracy of the estimated equal-loudness contour becomes poorer as the "true" underlying equal-loudness contour contains increasingly greater local details. When the complexity of the "true" equal-loudness contour is low (e.g., $k_{true} = 3$), the rms error increases gradually as the value of k_{IELC} increases. This is because the modeled contour with too many anchor points may cause overfitting. On the other hand, when the complexity of the "true"

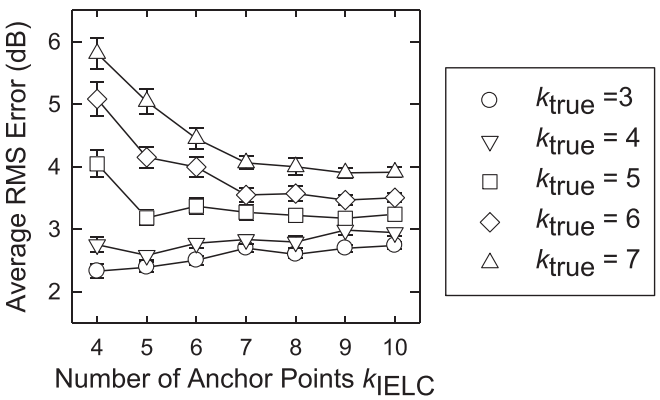

FIG. 8. The average rms error between the "true" simulated equal-loudness contour and the estimated contour via the IELC procedure. Results are shown as functions of k_{IELC}, i.e., the number of anchor points implemented for the IELC procedure. Various symbols indicate different value of k_{true}, a parameter that determines the complexity of the "true" equal-loudness contour (see text for details). Error bars indicate ± one standard error across 100 simulated listeners.

equal-loudness contour is high (e.g., $k_{true} = 7$), the rms error decreases rapidly as the value of k_{IELC} increases. This is because the modeled contour with too few anchor points may be insufficient in capturing the shape of the "true" equal-loudness contour. Therefore, the optimal choice of k_{IELC} in the IELC procedure requires knowledge on the complexity of the equal-loudness contour to be estimated. Too many or too few anchor points would undermine the accuracy of the estimated equal-loudness contour.

Besides the smoothness of the equal-loudness contour, another assumption of the IELC procedure is that the equal-loudness contour is static and does not change over time. However, it is possible that the listeners, when unfamiliar with the loudness comparison task, need to gradually establish their listening strategies and decision criteria. The time required for this familiarization process may be on the same order of an IELC run or even longer. In the current study, the first two IELC runs were conducted when the listeners were naive to the experimental task, while the third IELC run was conducted after more than 2000 trials of testing. Therefore, the fact that the first two IELC runs led to larger rms errors may be due to the lack of practice. It is still not clear, based on the current results alone, what the type or duration of practice is needed for a listener to establish stable listening strategies and criteria. Future studies are warranted.

B. Performance of the IML procedure

The second approach to improve the estimation of the equal-loudness contours involves implementing a Bayesian adaptive procedure (i.e., the UML procedure) in an interleaved fashion. This results in the development of the IML procedure. Similar to the ISC procedure, the IML procedure estimates the PSEs one frequency at a time. Since the IML procedure iteratively optimizes the stimulus selection according to the responses collected from previous trials, it has the potential of being more efficient. However, neither the test-retest reliability nor the accuracy of the IML procedure has been shown to be superior compared to the ISC procedure. Therefore, it seems that the IML procedure offers limited practical advantage over the ISC procedure.

The current study is not the first to implement Bayesian or maximum-likelihood procedures for the estimation of the equal-loudness contours. For example, Takeshima *et al.* (2001) used the maximum-likelihood procedure (Hall, 1968) to estimate the PSE in the equal-loudness contour. However, in that study, the maximum-likelihood procedure was implemented without interleaving tracks. The authors pointed out that since the maximum-likelihood procedure sampled stimuli at the PSE, listeners were often asked to compare test and standard stimuli of nearly equal loudness. To address this issue, an adaptive procedure, i.e. the randomized maximum-likelihood sequential (RMLS) procedure was proposed, in which the stimulus was set to the interim estimate of the PSE plus a random perturbation spanning from −6 to 6 dB. The addition of the random perturbation allowed for the sampling of points on the psychometric function with probabilities higher or lower than 50%. The IML procedure presented in the current study was similar to the RMLS but sampled high- and low-probability points (i.e., 29 and 71%) on the psychometric function using two interleaved tracks.

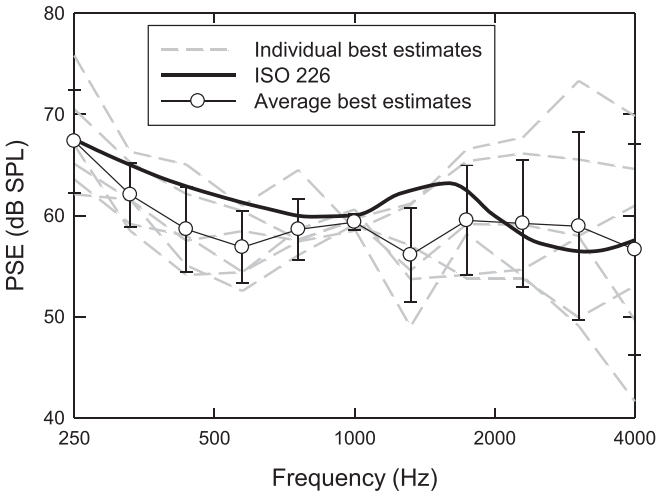

FIG. 9. The best estimates of the equal-loudness contours for the six listeners (dashed curves) together with the equal-loudness contour specified in ISO 226 (solid curve). Error bars indicate ± one standard deviation.

C. Individual differences in the equal-loudness contours

The average equal-loudness contour for normal-hearing listeners has been extensively studied (Fletcher and Munson, 1933; Robinson and Dadson, 1956; Watanabe and Møller, 1990; Takeshima *et al.*, 1997; Takeshima *et al.*, 2001; Suzuki and Takeshima, 2004) and standardized in ISO 226 (ISO, 2003). Figure 9 plots the best estimates of the equal-loudness contour for the six listeners (dashed curves) and their average (circles). The equal-loudness contour specified in ISO 226 (solid curve) is also shown as a reference. Although the equal-loudness contour from ISO 226 was within the range of individual best estimates from the current study, it exhibited certain differences from the average of the best estimates. It is worth pointing out that the equal-loudness contour from ISO 226 is for the binaural free field condition, while the equal-loudness contours from the current study were measured monaurally using a headphone. The headphones used in the current study, i.e., Sennheiser HD 280 Pro, were designed to meet the diffuse-field requirement for studio headphones (ITU-R BS.708, 1990). Therefore, the difference between the equal-loudness contours from the current study and ISO 226 may be contributed to (1) whether the frequency response of the headphones matched the free-field calibration target, and (2) the differences in the equal-loudness contours between the free-field and diffuse-field conditions.

Substantial individual differences were observed across the six listeners included in the current study; the standard deviations ranged between 0.8 and 10.4 dB across the 11 test frequencies (with a mean of 5.5 dB). The dependence of the individual differences on test frequency is in general agreement with previous studies. It has been shown that the standard deviation of the PSEs is the smallest near the frequency of the standard stimulus (i.e., 1 kHz in the current study) and

is the largest near the lowest and highest test frequencies (e.g., Reckhardt *et al.*, 1999). However, the previously reported standard deviations within the range of 250 and 4000 Hz are typically lower than 5 dB, while the standard deviations from the current study are larger (see the error bars in Fig. 9), especially for the highest test frequencies. The largest individual differences were observed at the highest two frequencies (9.3 dB at 3031 Hz and 10.4 dB at 4000 Hz). The sources of these large individual differences are not clear, but it seems that they are not dominated by intra-subject variabilities across various test runs. This is because satisfactory test-retest reliability was demonstrated in all three test procedures.

VI. SUMMARY

A Bayesian adaptive procedure, the IELC procedure, for the efficient estimation of equal-loudness contour was described in the current study. The IELC procedure with 200 loudness-comparison trials demonstrated excellent test-retest reliability, even when the test and retest were conducted at least a week apart. The test-retest reliability of the IELC procedure was comparable to that of the ISC and IML procedures with approximately 500 trials. Comparing to a "best estimate" of the equal-loudness contour based on the pooled data across all procedures and Exps. 1 and 2, the ISC and IML procedures showed similar accuracies. The IELC runs conducted when the listeners were naive to the loudness comparison task showed an accuracy that was poorer than the ISC and IML procedures, while the IELC runs conducted when the listeners were experienced with the task showed a comparable accuracy to the ISC and IML procedures.

ACKNOWLEDGMENTS

This work was supported by NIH Grant No. R21 DC013406 (Co-PIs: V. M. Richards and Y. Shen) and by the Training for Research and Academic Careers in Communication Sciences (TRACCS) program offered by the Department of Speech and Hearing Sciences at Indiana University.

[1]Kalman filtering is able to capture the effects of both observational noise and process noise. The observational noise is the noise caused by errors in the measurement, while the process noise is the noise inherent to the state variable. For the current application, we only included the observational noise in the model, since the characteristics of the process noise was hard to determine before data collection. Introducing the process noise to the model would reduce the effect of individual observations. This would cause the model parameters to converge slower; however, it would reduce the adverse influences of occasional aberrant responses.

[2]Instead of randomly sampling the test frequency, the IELC procedure can be implemented to adaptively select the test frequency on a trial-by-trial basis. Pilot experiments suggest that the adaptive sampling of frequency offers little benefit. As a result, random sampling was used with the IELC procedure in the current study.

Doire, C. S., Brookes, M., and Naylor, P. A. (**2017**). "Robust and efficient Bayesian adaptive psychometric function estimation," J. Acoust. Soc. Am. **141**, 2501–2512.

Epp, B., Verhey, J. L., and Mauermann, M. (**2010**). "Modeling cochlear dynamics: Interrelation between cochlea mechanics and psychoacoustics," J. Acoust. Soc. Am. **128**, 1870–1883.

Fletcher, H., and Munson, W. A. (**1933**). "Loudness, its definition, measurement and calculation," Bell Labs Tech. J. **12**, 377–430.

Florentine, M., Buus, S. R., and Poulsen, T. (**1996**). "Temporal integration of loudness as a function of level," J. Acoust. Soc. Am. **99**, 1633–1644.

Fritsch, F. N., and Carlson, R. E. (**1980**). "Monotone piecewise cubic interpolation," SIAM J. Numer. Anal. **17**, 238–246.

Hall, J. L. (**1968**). "Maximum-likelihood sequential procedure for estimation of psychometric functions," J. Acoust. Soc. Am. **44**, 370–370.

Hatzfeld, C., Kupnik, M., and Werthschützky, R. (**2015**). "Performance simulation of unforced choice paradigms in parametric psychometric procedures," in *Proceedings of the 2015 IEEE World Haptics Conference (WHC)*, June 22–26, Chicago, IL, pp. 475–481.

International Organization for Standardization (**2003**). ISO 226, *Acoustics-Normal Equal-Loudness Contours* (International Organization for Standardization, Geneva, Switzerland).

International Telecommunication Union (**1990**). ITU-R BS. 708, *Determination of the Electro-Acoustical Properties of Studio Monitor Headphones* (International Telecommunications Union, Geneva, Switzerland).

Jesteadt, W. (**1980**). "An adaptive procedure for subjective judgments," Atten. Percept. Psychophys. **28**, 85–88.

King-Smith, P., and Rose, D. (**1997**). "Principles of an adaptive method for measuring the slope of the psychometric function," Vis. Res. **37**, 1595–1604.

Kontsevich, L. L., and Tyler, C. W. (**1999**). "Bayesian adaptive estimation of psychometric slope and threshold," Vis. Res. **39**, 2729–2737.

Levitt, H. C. C. H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Marks, L. E. (**1988**). "Magnitude estimation and sensory matching," Percept. Psychophys. **43**, 511–525.

Marks, L. E. (**1993**). "Contextual processing of multidimensional and unidimensional auditory stimuli," J. Exp. Psychol. Hum. Percept. Perform. **19**, 227–249.

Marks, L. E., and Florentine, M. (**2011**). "Measurement of loudness, Part I: Methods, problems, and pitfalls," in *Loudness* (Springer, New York), pp. 17–56.

Mauermann, M., Long, G. R., and Kollmeier, B. (**2004**). "Fine structure of hearing threshold and loudness perception," J. Acoust. Soc. Am. **116**, 1066–1080.

Reckhardt, C., Mellert, V., and Kollmeier, B. (**1999**). "Factors influencing equal-loudness level contours," in *Psychophysics, Physiology and Models of Hearing* (World Scientific, Singapore), pp. 113–116.

Robinson, D. W., and Dadson, R. S. (**1956**). "A re-determination of the equal-loudness relations for pure tones," Br. J. Appl. Sci. **7**, 166–181.

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. Meth. **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.

Shen, Y., Sivakumar, R., and Richards, V. M. (**2014**). "Rapid estimation of high-parameter auditory-filter shapes," J. Acoust. Soc. Am. **136**, 1857–1868.

Suzuki, Y., and Takeshima, H. (**2004**). "Equal-loudness-level contours for pure tones," J. Acoust. Soc. Am. **116**, 918–933.

Takeshima, H., Suzuki, Y., Fujii, H., Kumagai, M., Ashihara, K., Fujimori, T., and Sone, T. (**2001**). "Equal-loudness contours measured by the randomized maximum likelihood sequential procedure," Acta Acust. united Ac. **87**, 389–399.

Takeshima, H., Suzuki, Y., Kumagai, M., Sone, T., Fujimori, T., and Miura, H. (**1997**). "Equal-loudness levels measured with the method of constant stimuli," J. Acoust. Soc. Japan **18**, 337–340.

Verhey, J. L., and Kollmeier, B. (**2002**). "Spectral loudness summation as a function of duration," J. Acoust. Soc. Am. **111**, 1349–1358.

Watanabe, T., and Møller, H. (**1990**). "Low frequency hearing thresholds in pressure field and in free field," J. Low Freq. Noise Vib. Active Control **9**, 106–115.

Watson, A. B., and Pelli, D. G. (**1983**). "QUEST: A Bayesian adaptive psychometric method," Percept. Psychophys. **33**, 113–120.

Zwicker, E., Flottorp, G., and Stevens, S. S. (**1957**). "Critical band width in loudness summation," J. Acoust. Soc. Am. **29**, 548–557.

Zwislocki, J. J. (**1969**). "Temporal summation of loudness: An analysis," J. Acoust. Soc. Am. **46**, 431–441.