

Hear-to-Hear: Uncertainty-Aware, Head-Trackable Interaction for Personalized Spatial Audio

Yu-Chen Tsai
National Taiwan University
Taipei, Taiwan

Lung-Pan Cheng
National Taiwan University
Taipei, Taiwan

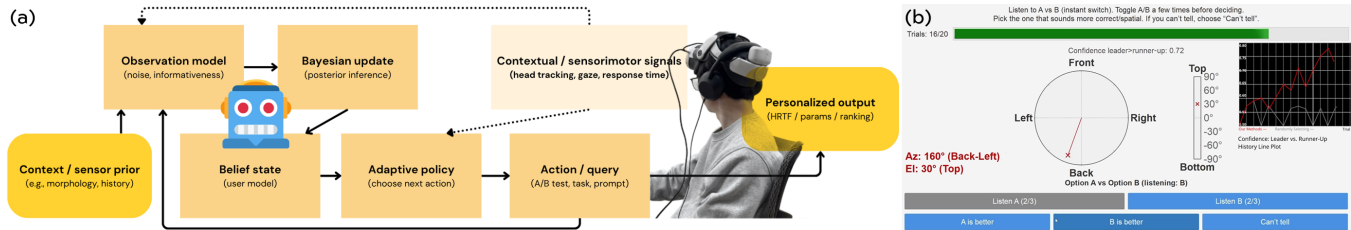


Figure 1: Real-time Bayesian human-in-the-loop personalization. (a) The framework uses a prior and observation model to interpret feedback, Bayesian updating to maintain a user belief state, and an adaptive policy to select the next query and produce a personalized configuration. (b) Pairwise selection interface for head-tracked HRTF personalization, with a confidence visualization shown for illustration only and not presented during the user study.

Abstract

Selecting a personalized head-related transfer function (HRTF) is essential for accurate spatial audio, but remains challenging due to substantial user variation and noisy subjective feedback. We present a real-time, uncertainty-aware perceptual interaction loop for HRTF personalization: after each paired comparison, the system updates a posterior belief over candidate HRTFs and immediately adapts the next query to maximize informativeness. Our system couples Bayesian preference learning with real-time head-tracked binaural rendering, enabling head motion to actively inform comparisons. In a controlled within-subject psychophysical study, uncertainty-aware interaction (with optional head tracking) provides directional posterior evidence for localization improvements—primarily suggestive reductions in azimuth error and front-back confusions concentrated in the front hemisphere—alongside participant-level heterogeneity and a partially mitigated elevation trade-off, rather than uniformly decisive effects across all metrics.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools.**

Keywords

Bayesian active preference learning, Uncertainty modeling, Head-related transfer function, Spatial audio, Personalization



This work is licensed under a Creative Commons Attribution 4.0 International License.
UIST '26, Detroit, MI, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2856-3/2026/11
<https://doi.org/10.1145/3830398.3830661>

ACM Reference Format:

Yu-Chen Tsai and Lung-Pan Cheng. 2026. Hear-to-Hear: Uncertainty-Aware, Head-Trackable Interaction for Personalized Spatial Audio. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26)*, November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3830398.3830661>

1 Introduction

Realistic spatial audio in stereo is achieved using binaural models that mimic human hearing. These systems use head-related transfer functions (HRTFs) to recreate sound direction, distance, and spatial stability [4, 12]. Individual ear, head, and torso morphology alters how sound reaches the ears and affects HRTF spatial localization accuracy. Because HRTFs vary substantially across individuals, population-average HRTFs can cause poor externalization and front-back confusions [21, 22, 30]. Measuring individualized HRTFs offers high fidelity but is yet impractical at scale [2], motivating scalable interactive personalization. The scalable approaches either (i) predict HRTFs from anthropometry/ear images [32] or (ii) ask users to select from a small database [20]; both are limited by model bias and by noisy, effortful judgments. Spatial perception is also shaped by self-motion, yet head tracking is rarely treated as a first-class signal for making preference elicitation more informative.

In this paper, we propose an approach to personalize HRTF using preference learning through head-contingent auditory cues. We model each listener as a latent preference profile over candidate HRTFs with an explicit posterior, which is the user state at time t ; adaptation is implemented as a personalized querying policy selecting the next paired comparison conditioned on that state.

At the highest level, this brings a *real-time Bayesian interaction loop* into perceptual personalization, where interaction is uncertainty-aware adaptive querying: the system maintains a posterior over candidates and chooses the next comparison to maximize informativeness under noisy human feedback, enabling sensing and

querying to co-adapt. The key is a *minimal-sensing interaction design* for perceptual personalization without instrumentation: no ear images, no anthropometry, and no personalized ground-truth HRTF are required; an ear-measure-based prior is an optional accelerator. In addition, we found that head tracking can act as an information amplifier (not a dependency), and it is achievable with consumer-grade hardware. It increases query efficiency by making comparisons more diagnostic through head-contingent cues. This makes HRTF personalization more scalable without requiring user ear images or raising privacy concerns.

We evaluate our approach in a controlled within-subject psychophysical study with a fixed-HRTF baseline (non-interactive) and uncertainty-aware adaptive querying (interactive), crossed with head tracking (on/off) and with/without an ear-measure-based prior (as optional acceleration). From a pool of 93 candidate HRTFs, a bounded budget of $T_{\max} = 20$ comparisons (≈ 6 minutes) yields directional posterior evidence for localization improvements concentrated in the front hemisphere (suggestive azimuth and front-back reductions), with participant-level heterogeneity and a partially mitigated elevation trade-off, relative to a fixed-HRTF baseline; we do not claim generalization beyond this pointing-task paradigm.

To improve reproducibility, we provide the data and the code for the prototype interaction interface used in the study.¹

2 Related Work

HRTF personalization can be achieved in several ways. These include physical measurement and simulation, as well as morphological prediction using anthropometric data or ear images. Other methods rely on interactive perceptual selection, where users judge candidate HRTFs. More recently, individualization based on machine learning has also emerged. We connect these approaches by framing personalization as human-in-the-loop inference, in which self-motion cues and uncertainty-aware interaction increase the information extracted from each judgment.

HRTF Measurement, Datasets, and User Selection. Individual HRTFs are measured in controlled setups or simulated from 3D geometry [1], but both are limited for professional use. Public datasets (CIPIC [2], HUTUBS [7], SADIE II [15]) enable standardized evaluation. Paired-comparison protocols can let user select from measured databases to outperform static baselines, but are time-consuming for large dataset and sensitive to noisy judgments [16].

Morphology-Based HRTF Individualization. Early studies match simple body measurements with existing HRTF databases [27]. More recently, researchers have employed deep neural networks to construct personalized HRTF directly [11], predict how individuals perceive sounds from different directions [31], and capture highly specific acoustic details [3]. Such methods reduce user burden and offer an efficient *prior*, but they are limited by dataset/model bias and the imperfect link between physical features and subjective preference, often missing idiosyncratic or perceptually important errors and thus motivating interactive refinement.

Head-Trackd Rendering. Dynamic listening improves spatial perception: head motion provides time-varying cues that reduce

confusions and support externalization and stability [6, 28]. Martin and Picinali [18] found that individualized HRTFs increased perceived realism in static listening, but enabling head movement altered the pattern of benefits for realism versus localization. Unlike prior work that often treats head tracking as a rendering enhancement, we treat head-contingent cues as an information source that can change *query efficiency* during personalization.

Preference Learning Under Uncertainty. Paired comparisons are often modeled using Thurstone or Bradley–Terry approaches [5, 26], with extensions like Davidson and Rao–Kupper to handle ties (“no preference”) [10, 24]. Bayesian methods, e.g. Gaussian process preference learning and TrueSkill-style models, are also widely used for learning from comparative feedback [8, 13]. Active preference learning then chooses which comparisons to show by maximizing informativeness (e.g., expected information gain) [14, 19], which we adopt to reduce interaction burden under noisy feedback.

Summary and Positioning. Building on scalable morphology-based priors, perceptual selection methods, and evidence that head motion shapes spatial hearing, we frame HRTF personalization as a standard user-adaptation loop: maintain a latent user state (preferences + uncertainty), update it from interaction, and adapt system behavior via query selection (personalized information gathering). This unifies head-contingent cues and uncertainty-aware interaction as first-class signals for efficient personalization while surfacing cold-start, heterogeneity, stability/drift, and experience-level outcomes alongside accuracy.

3 Bayesian User Model and Adaptation

Users go through a Bayesian adaptive pairwise-comparison stage (Fig. 1) used to select a single “best” HRTF subject from a finite set (e.g., 93 HUTUBS HRTFs, or 94 if the user’s own ear-measured HRTF is included). The listener performs direct comparisons between two HRTF binaural renderings; a TrueSkill-style probit model maintains a latent utility and uncertainty per subject and is updated after each comparison. Pair selection is adaptive: an initial mini-tournament over a small diverse subset is followed by a main phase that chooses the next pair to maximize expected information gain (EIG) about the identity of the best HRTF. This stage ends when the current leader is very likely to beat the runner-up, meaning the posterior probability exceeds a confidence threshold (e.g., 0.95). The chosen best is then used in the localization test.

3.1 The personalization process

Each HRTF item $i \in 1, \dots, n$ is associated with a latent preference strength (utility) u_i , which is assumed to follow a normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$, where μ_i and σ_i^2 denote the posterior mean and posterior variance of the utility, respectively. When comparing i vs. j , the observed performance difference is $d_{ij} = (u_i - u_j) + \eta$ where $\eta \sim \mathcal{N}(0, 2\beta_{\text{eff}}^2)$, and β_{eff} is the effective trial noise. The outcome is determined by the sign of d_{ij} relative to a draw margin $\varepsilon \geq 0$: i wins if $d_{ij} > \varepsilon$, j wins if $d_{ij} < -\varepsilon$, and it is a draw (“Can’t tell”) if $-\varepsilon \leq d_{ij} \leq \varepsilon$. The posterior over $(\mu_i, \sigma_i, \mu_j, \sigma_j)$ is updated after each comparison using closed-form probit updates.

¹<https://github.com/yuchen-tsai/Hear-to-Hear>

(1) **Onboarding:** A mini-tournament is run on a diverse subset of $k = 5$ candidate indices (selected via k -means on HRTF PCA coordinates, with the ear-measure-based predicted HRTF optionally appended). First, HRTF_1 and HRTF_2 are compared; the winner then faces HRTF_3 , and so on, taking $k - 1$ trials to seed the posteriors.

(2) **Update Equations:** Latent utilities are updated using a TrueSkill-style probit model. Without loss of generality, let i denote the winner and j the loser. The update is driven by the standardized difference $d = (\mu_i - \mu_j)/c$, where $c^2 = \sigma_i^2 + \sigma_j^2 + 2\beta_{\text{eff}}^2$ and draw margin ε . Define $t = (\mu_i - \mu_j - \varepsilon)/c$, $v = \phi(t)/\Phi(t)$, and $w = v(v + t)$, where ϕ and Φ are the standard normal PDF and CDF. The posterior parameters are then updated as $\mu'_k = \mu_k + s_k \sigma_k^2 v/c$ and $\sigma_k'^2 = \sigma_k^2 \max(\kappa, 1 - \sigma_k^2 w/c^2)$ for $k \in \{i, j\}$, where $s_i = +1$, $s_j = -1$, and the variance floor is $\kappa = 10^{-9}$.

If a draw is observed, we define $a = (-\varepsilon/c) - d$ and $b = (\varepsilon/c) - d$. The likelihood-based updates use:

$$v = \frac{\phi(a) - \phi(b)}{\Phi(b) - \Phi(a)}, \quad w = v^2 + \frac{a\phi(a) - b\phi(b)}{\Phi(b) - \Phi(a)}.$$

Applying v and w to the same mean and variance update equations pulls $\mu_i - \mu_j$ toward 0 (indifference) and reduces uncertainty, avoiding overconfidence when listeners are unsure.

We interpret query selection as *adaptive system behavior* conditioned on the user state. To handle lapse rates λ and early-trial uncertainty, the update uses an effective noise β_{eff} :

$$\beta_{\text{eff}} = \underbrace{\beta \cdot \left(1 + \alpha_{\text{anneal}} e^{-t/\tau_{\text{anneal}}}\right)}_{\text{early-trial annealing}} \cdot \underbrace{\left(1 + \alpha_{\text{rt}} e^{-r/\tau_{\text{rt}}}\right)}_{\text{response-time (optional)}} \cdot \underbrace{\frac{1}{1 - \lambda}}_{\text{lapse inflation}},$$

where t is the current trial, r is response time in seconds.

(3) **EIG:** To select the next pair (i, j) from the candidate set (leader, runner-up, top- K_μ means, and top- K_σ uncertainties), the system maximizes a Monte Carlo approximation of the EIG about the best HRTF W , with exploration probability $\varepsilon_{\text{explore}} = 0.06$. The prior winner entropy is estimated using 450 Monte Carlo samples, and each outcome in the EIG calculation uses 300 samples.

The system schedules queries by choosing the next pair (i, j) to maximize Expected Information Gain (EIG) regarding which candidate is the true best $W = \arg \max_k u_k$:

$$\text{EIG}(i, j) = H(W) - \mathbb{E}_y[H(W | y)],$$

where $y \in \{\text{win, loss, draw}\}$ is the outcome.

(4) **Stopping Criteria:** The stage exits when the probability $P(\text{leader beats runner-up}) \geq p_{\text{stop}}$ and the leader wins a confirmation trial against the runner-up. To balance information gain, participant burden, and session length in a within-subject design, all interactive conditions were evaluated under a matched interaction budget of $T_{\text{max}} = 20$ comparisons; this cap reflects a fatigue compromise rather than a convergence guarantee.

Hyperparameters used in our study are summarized in Table 1.

4 Evaluation

The goal of this evaluation is to test an interaction paradigm in which *interaction is uncertainty-aware adaptive querying*: the system uses in-situ perceptual feedback to update a posterior over candidate HRTFs and chooses the next comparison to be maximally informative under noisy judgments. We compare this to a

Table 1: Hyperparameters of the Bayesian personalization model.

Symbol	Value	Description
σ_0	1.0	Initial utility uncertainty.
β	0.6	Base comparison noise.
ε	0.25	Draw margin (half-width of draw region).
λ	0.06	Lapse rate (random choice probability).
α_{anneal}	1.0	Early-trial annealing scale.
τ_{anneal}	8.0	Annealing time constant (trials).
α_{rt}	0.8	Response-time noise scale.
τ_{rt}	0.7	RT time constant (seconds).
K_μ, K_σ	7, 7	Candidate pool size filters.
p_{stop}	0.95	Confidence threshold for stopping.
T_{max}	20	Interaction trial budget cap (trials).

non-interactive fixed-HRTF baseline, and additionally test whether head tracking improves diagnostic comparisons.

4.1 Participants and Apparatus

We recruited $N = 16$ participants (9 males, 7 females; aged 20–29, $M = 22.94$, $SD = 3.05$), all reporting normal hearing and no known vestibular disorders. No prior experience with spatial audio. The study followed a Latin-square counterbalanced within-subjects design. Participants were assigned to eight Latin square conditions in randomized recruitment order so each condition included two participants. All participants completed the full study.

We selected a pool of 93 HRTFs from the HUTUBS database [7], chosen to span substantial anthropometric diversity and excluding those without ear anthropometry data. Our apparatus consists of consumer grade hardware. Audio was presented binaurally over circumaural *Sony WH-1000XM6* headphones in a quiet room. To avoid latency, jitter, and DSP coloration that degrade spatial cues, the headphones were used in powered-off passive (wired) mode. Head tracking was achieved at ≈ 70 Hz using a *Meta Quest 3* VR head-mounted display via *Quest Link*. A *Godot Engine* application on a Windows 11 PC streamed head pose to the experiment software through a lightweight Python UDP server, and participants sat on a swivel chair to allow natural, unconstrained rotation. The VR headset can be removed during deployment, since many consumer-grade headphones already track head movement, though they do not provide access to raw data. System responsiveness met real-time requirements, with end-to-end query updates completing in under 1 second and no perceptible lag in head-contingent rendering.

The total session duration was approximately 150 minutes, comprising: ear measurements (≈ 15 min), four rounds of 20 pairwise comparisons (≈ 6 min per round), eight localization tests (≈ 13 min each), and feedback forms (≈ 7 min). Stimuli playback alternated between directions ($20^\circ, -30^\circ$) and ($160^\circ, 30^\circ$) for Group 1, and ($340^\circ, -30^\circ$) and ($200^\circ, 30^\circ$) for Group 2. Playback actions were limited to 3 per stimulus to manage fatigue.

4.2 Experimental Design and Conditions

The experiment employed a 2 (Head Tracking) $\times$ 4 (HRTF Method) factorial design, resulting in eight experimental conditions²:

²We use the abbreviations (g, gh, e, ...) for conditions hereafter.

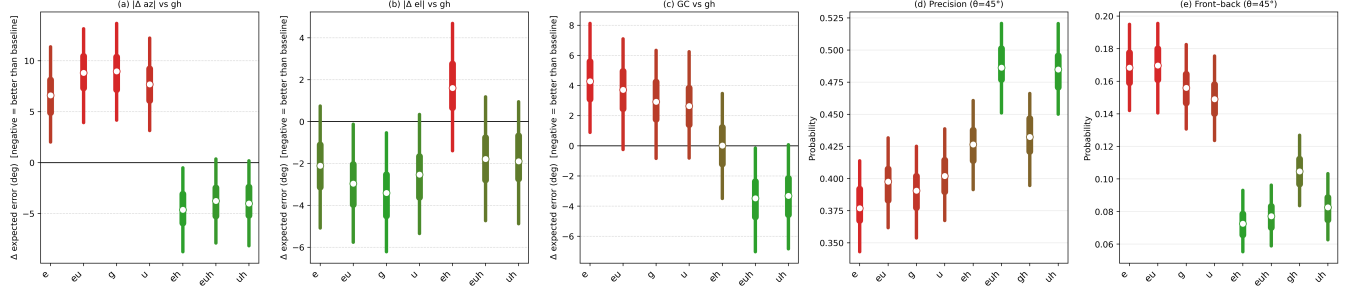


Figure 2: Posterior Bayesian summaries vs. baseline gh. (a) Absolute azimuth error contrasts ($|\Delta\Phi|$; lower Δ is better). (b) Absolute elevation error contrasts ($|\Delta\phi|$; lower Δ is better). (c) Great-circle error contrasts (lower Δ is better). (d) Precision category proportions at $\theta = 45^\circ$ (higher is better). (e) Front-back confusion proportions at $\theta = 45^\circ$ (lower is better). Points are posterior medians; thick and thin bars denote 50% and 95% highest-density intervals (HDIs).

- g: Generic / Tracking OFF
- gh: Generic / Tracking ON
- e: Ear-measured / Tracking OFF
- eh: Ear-measured / Tracking ON
- eu: Ear-measured + Preference / Tracking OFF
- euh: Ear-measured + Preference / Tracking ON
- u: Preference-only / Tracking OFF
- uh: Preference-only / Tracking ON

In conditions with preference learning, comparisons were scheduled and posteriors were updated using the Bayesian user model and adaptive pair-selection policy described in Section 3.

This design separates (i) a non-interactive fixed-HRTF baseline from (ii) uncertainty-aware adaptive querying, while treating ear-measure-based individualization as an optional prior (acceleration mechanism) rather than as a required input modality. It also isolates head tracking as an information amplifier (ON/OFF) rather than a dependency of the interaction loop.

4.2.1 Head Tracking Manipulation. In the ON condition, audio updated in real time based on participants' head movements to provide motion-contingent cues, whereas in the OFF condition, sound rendering remained static without coupling to head motion.

4.2.2 HRTF Method Conditions. **Generic Baseline:** A fixed, non-personalized, population-average HRTF (specifically, the pp31 HRTF from the HUTUBS dataset, representing the center of its Principal Component Analysis space) was used throughout, with no interaction or adaptation. **Ear-measured Individualized:** We collected ear measurements from each participant (approximately 15 minutes each). A single-shot HRTF prediction was obtained from ear measurements by performing inference on the HRTF personalization model of Wang et al. [29]. We retrained the model using a reduced set of anthropometric variables (dropping 14 of the original 28 inputs). In our leave-one-out cross-validation (LOOCV) baseline using their original implementation, we obtained a log-spectral distance (LSD) of 4.155 dB, whereas our restricted-input model achieved an LSD of 4.099 dB (validation MSE: 0.896 vs. 0.875), representing comparable reconstruction quality under significantly reduced user input requirements. **Preference-only:** Interactive, uncertainty-aware preference learning initialized from the generic baseline HRTF, with no morphological prior. The 20 pairwise comparisons take about 6 minutes for participants. **Ear-measured +**

Preference Learning: Interactive, uncertainty-aware preference learning initialized from the ear-measure-based HRTF.

4.3 Localization Tasks and Metrics

This stage assesses spatial localization accuracy under the 8 conditions described above (Section 4.2). For the four preference-learning conditions, the localization test was administered *after* the pairwise selection stage using an interface designed for the task (Fig. 1b) and used the single winner HRTF selected for that participant in that condition; for the non-interactive conditions (Generic and Ear-measured), we used the corresponding fixed HRTF throughout the localization trials. Participants are not informed about how their actions in the pairwise selection stage affect the localization test.

Each trial presents a mono speech stimulus (48 kHz) from TSP Speech, with the speaker and recording randomly selected while preventing immediate repetition [25]. Rendering details (head-relative coordinates, HRTF interpolation, convolution) use real-time head pose to update head-relative direction, selecting the nearest candidate HRTF impulse response pair and applying overlap-add convolution. Trials were drawn from a predefined set of spatial directions spanning five elevations ($+60^\circ$, $+30^\circ$, 0° , -30° , -60°) and azimuth increments of 40° (0° – 320°), yielding a total of 45 unique target directions. The trial sequence was randomized once at initialization to minimize systematic ordering effects. Participants listen to spatialized audio from fixed azimuth–elevation positions and report the perceived sound direction via an interactive localization interface (Fig. 3). To ensure data integrity, a monitoring routine continuously assessed head-pose motion during active recording.

To analyze localization accuracy, we define five primary metrics: (1) **Great-circle error:** The angular error $\delta(T, R) = \arccos(\mathbf{u}_T \cdot \mathbf{u}_R)$ (in degrees) between the target unit vector $\mathbf{u}_T$ and response unit vector $\mathbf{u}_R$. (2) **Azimuth and elevation error:** The signed azimuth error is $\Delta_{az} = ((r_{az} - t_{az} + 180^\circ) \bmod 360^\circ) - 180^\circ$, which is wrapped to $[-180^\circ, 180^\circ]$. The elevation error is $\Delta_{el} = r_{el} - t_{el}$, where $\Delta_{el} \in [-180^\circ, 180^\circ]$. (3) **Precision:** Classified when the great-circle error is small, $\delta(T, R) < 45^\circ$. (4) **Front-back classification:** Classified if the trial does not meet precision but the response is close to the mirrored target direction across the frontal plane, i.e., $\delta((t'_{az}, t_{el}), R) < 45^\circ$, where $t'_{az} = (180^\circ - t_{az}) \bmod 360^\circ$. (5) Lateral and polar errors were measured using the methodology described by Poirier-Quinot et al. [23].

4.4 Results

We summarize contrast uncertainty with Bayesian Highest Density Intervals (HDIs) and directional probabilities (e.g., $P(\Delta < 0)$); intervals overlapping zero are interpreted as suggestive rather than decisive (metric definitions in Section 4.3).

In Fig. 2a, relative to the strongest baseline condition gh, posterior distributions of contrast effects Δ show suggestive azimuth error $|\Delta\phi|$ reductions under personalization: median contrasts show -4.02° for uh-gh (95% HDI/CrI: $[-8.17^\circ, 0.18^\circ]$, $P(\Delta < 0) = 0.972$; 8/16 users improved) and -3.77° for euh-gh (CrI: $[-7.89^\circ, 0.36^\circ]$, $P(\Delta < 0) = 0.965$). Because the credible intervals overlap zero slightly, we interpret these as directional or posterior evidence rather than uniformly decisive effects.

For absolute elevation error $|\Delta\phi|$ (Fig. 2b), condition g outperforms gh by -3.41° (95% CrI: $[-6.22^\circ, -0.54^\circ]$, $P(\Delta < 0) = 0.991$), providing posterior evidence for tracking-induced elevation degradation (CrI excludes zero). Personalization shows a suggestive mitigation of this trade-off relative to gh, with median contrasts of -1.90° for uh (CrI: $[-4.88^\circ, 0.95^\circ]$, $P(\Delta < 0) \approx 0.895$) and -1.79° for euh (CrI: $[-4.73^\circ, 1.18^\circ]$, $P(\Delta < 0) \approx 0.882$); since these intervals include zero, the results are directional rather than decisive. Descriptive analyses localize tracking-induced elevation degradation to lower target elevations ($+5.44^\circ$ for gh-g, $+3.45^\circ$ for uh-u) and horizontal targets ($+4.87^\circ$ / $+2.45^\circ$), whereas upper targets show no degradation and instead suggest improvement under personalization by -2.87° for uh-u. This elevation trade-off is participant-specific: 10/16 participants show worse elevation under uh relative to u (e.g., u_{41} degraded by $+21.1^\circ$ using candidate pp_{10}), while others improve (e.g., u_{37} improved by -18.9° using candidate pp_{12}).

For great-circle error (Fig. 2c), personalization shows a suggestive reduction relative to gh for uh-gh and posterior evidence for reduction for euh-gh. Median contrasts are -3.32° for uh-gh (95% CrI: $[-6.83^\circ, 0.07^\circ]$, $P(\Delta < 0) = 0.971$; 11/16 participants improved; CrI includes zero) and -3.48° for euh-gh (CrI: $[-7.02^\circ, -0.15^\circ]$, $P(\Delta < 0) = 0.976$; CrI excludes zero). Descriptive sector analyses (not posterior contrasts) suggest that changes concentrate in the front hemisphere (target azimuths within $[-90^\circ, 90^\circ]$, where baseline errors are high): mean front GC error decreases from 61.32° (gh) to 55.52° (uh, -5.80°) and 55.04° (euh, -6.28°); mean front azimuth error decreases from 44.05° (gh) to 38.78° (uh, -5.27°) and 37.71° (euh, -6.34°); and front-back confusions decrease from 14.0% (gh) to 9.8% (uh, -4.2 pp) and 9.2% (euh, -4.8 pp). Conversely, rear-hemisphere mean errors remain similar (rear GC is 52.02° for gh versus 51.34° for uh).

Participant-level outcomes are heterogeneous: 11/16 participants show lower great-circle error under uh-gh, 8/16 show lower azimuth error, and elevation remains mixed (10/16 show worse elevation under uh relative to u; 13/16 show lower elevation under uh-gh). Taken together, these analyses provide directional posterior evidence for front-hemisphere improvements in azimuth and front-back disambiguation, posterior evidence for great-circle reduction under euh-gh (and a suggestive reduction under uh-gh), and a partially mitigated, participant-specific elevation trade-off.

Looking at the posterior category proportions, Precision ($GC < 45^\circ$, Fig. 2d) shows posterior evidence of improvement, increasing from 43.2% (gh) to 48.5% (uh, $Pr(\Delta > 0) = 0.976$) and 48.6% (euh,

$Pr(\Delta > 0) = 0.984$). The proportion of Front-back confusions (Fig. 2e) shows directional posterior evidence of reduction, decreasing from 10.5% (gh) to 8.3% (uh, $Pr(\Delta < 0) = 0.927$) and 7.7% (euh, $Pr(\Delta < 0) = 0.971$). Polar error shows a suggestive mean reduction of -5.29° for uh-gh (67.20° vs. 72.49°), while lateral error remains stable (19.69° vs. 19.93°). We conducted a sensitivity analysis across thresholds $\theta = 25^\circ, 35^\circ, 45^\circ, 55^\circ$, and 65° , showing that these findings are qualitatively robust. The classification of $\theta = 45^\circ$ as “precision” follows prior work [18].

4.5 Discussion

Our results suggest a perceptually plausible pattern rather than uniform improvement: uncertainty-aware personalization provides directional posterior evidence for front-hemisphere improvements—suggestive azimuth reductions and posterior evidence for reduced front-back confusions—with posterior evidence for great-circle reduction under euh-gh and a suggestive reduction under uh-gh. Elevation exhibits a trade-off: head tracking degrades elevation versus static rendering (posterior evidence; CrI excludes zero), and personalization shows a suggestive mitigation of this loss; participant-level outcomes remain heterogeneous. We treat any spectral explanation as a hypothesis rather than a confirmed mechanism, given that speech stimuli, despite providing broadband and time-varying localization cues, may not adequately represent the high-frequency spectral cues required for accurate elevation perception.

The budget of $T_{\max} = 20$ was chosen as a fatigue compromise rather than a convergence guarantee. Replaying trial logs indicates that preferences continue to refine late in the session: the leader at trial T matches the final winner in only 29.7% of cases at $T = 10$, and 50.0% at $T = 15$. Despite this tight budget, synthetic simulations (runs=200, $T = 20$) show modest but consistent gains under EIG selection: it selects a top-10 HRTF in 31.0% of cases (vs. 24.5% for random selection) with a lower mean rank (26.66 vs. 32.46) and lower regret (1.81 vs. 1.90). EIG also selects the absolute best HRTF in 8.5% of cases (vs. 6.5% for random). These findings support EIG as a useful, budget-aware querying strategy. Our approach supports *personalization without instrumentation*: it can run purely from in-situ feedback, without ear images or anthropometry. Head tracking optionally amplifies information, making queries more diagnostic, and the uh-gh comparison allows the personalization effect to be evaluated independently of head tracking.

The proposed framework is complementary to recent estimation and personalization approaches rather than a replacement for them. For example, a learned HRTF reconstruction pipeline taking pinna images as input such as PRTFNet [17] could be used to generate a stronger individualized prior or candidate ranking, after which our preference-learning loop could refine the result using perceptual feedback. The LAP24 challenge [9] could help future work study how robust our preference-learning framework is across harmonized, multi-dataset candidates and how effectively preference learning can be used to upsample sparse HRTF measurements.

5 Limitations and Future Work

Our evaluation is limited by ecological validity, and our claims are restricted to controlled psychophysical evidence. Pointing tasks do not automatically generalize to real-world VR search, gaming, or

physical pointing. Future work should evaluate ecological generalization in immersive or task-based settings (e.g., VR search, gaming, or naturalistic pointing). Our cohort comprised $N = 16$ young adults (20–29 years) selecting from a finite database of 93 candidate HRTFs over a single short session. All participants were novice listeners, and closed-back headphones were used in passive wired mode to standardize consumer usage; these factors were shared across conditions but may limit demographic and ecological generalizability. While mild fatigue was reported by some participants, this can be mitigated in future work by integrating the preference-learning loop into interactive onboarding sequences (such as game tutorials) rather than sterile pairwise comparisons.

Our method reduces, but does not eliminate, the noisy nature of subjective auditory preferences. While ear-measure-based individualization provides a prior, its effectiveness depends on the accuracy and coverage of the morphological model. Hyperparameters ($T_{\max}$, β , and others) were derived from pilot tuning rather than optimization, and outcomes remain sensitive to these settings.

Our system performs database selection rather than predicting ground-truth HRTFs. Because several contrasts are suggestive rather than decisive (CrIs overlapping zero), and because participant-level outcomes are heterogeneous, future work should explore multi-objective acquisition, elevation-aware stimuli, and sector-wise evaluation beyond front-hemisphere analyses. Future work should also validate across diverse demographics and adapt this sensorimotor personalization loop (Fig. 1) to other perceptual modalities (e.g., haptics, accessibility controls, or AR/VR interfaces).

6 Conclusion

This work demonstrates that real-time, uncertainty-aware adaptive querying can reduce perceptual personalization to a few minutes (≈ 6 minutes; 20 comparisons) while yielding directional posterior evidence for downstream localization improvements in a controlled pointing task—primarily suggestive front-hemisphere reductions in azimuth error and front-back confusions, with participant-level heterogeneity and a partially mitigated elevation trade-off—rather than uniformly decisive effects across all spatial metrics. By formulating HRTF personalization as Bayesian inference with active experimental design, we make adaptive querying itself the interaction paradigm, and we show how head motion (via optional head tracking) can serve as an information channel that amplifies the informativeness of in-situ perceptual feedback within this psychophysical evaluation setting.

Acknowledgments

This work was supported by National Taiwan University (NTU-CC-115L890902). We thank Prof. Ja-Ling Wu for insightful feedback during the development of this course project.

References

- [1] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson. 2002. Approximating the Head-Related Transfer Function Using Simple Geometric Models of the Head and Torso. *The Journal of the Acoustical Society of America* 112, 5 (2002), 2053–2064.
- [2] V. Ralph Algazi, Richard O. Duda, Dennis M. Thompson, and Carlos Avendaño. 2001. The CIPIC HRTF Database. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 99–102.
- [3] Lior Arbel, Ishwarya Ananthabhotla, Zamir Ben-Hur, David Lou Alon, and Boaz Rafaely. 2024. On HRTF Notch Frequency Prediction using Anthropometric Features and Neural Networks. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 816–820. doi:10.1109/ICASSP48485.2024.10447610
- [4] Jens Blauert. 1996. *Spatial Hearing: The Psychophysics of Human Sound Localization* (revised edition ed.). The MIT Press.
- [5] Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 3–4 (1952), 324–345. doi:10.1093/biomet/39.3-4.324
- [6] W. Owen Brimijoin, Andrew W. Boyd, and Michael A. Akeroyd. 2013. The Contribution of Head Movement to the Externalization and Internalization of Sounds. *PLoS ONE* 8, 12 (2013), e83068.
- [7] Fabian Brinkmann, Alexander Lindau, and Stefan Weinzierl. 2019. HUTUBS: A Database of Head-Related Transfer Functions and Anthropometric Features. TU Berlin, DepositOnce. doi:10.14279/depositonce-8487 HRTF database and documentation.
- [8] Wei Chu and Zoubin Ghahramani. 2005. Preference Learning with Gaussian Processes. In *Proceedings of the International Conference on Machine Learning (ICML)*. 137–144.
- [9] Rapolas Daugintis, Roberto Barumerli, Michele Geronazzo, Johan Pauwels, Lorenzo Picinali, and Katarina C. Poole. 2025. Listener Acoustic Personalization Challenge—LAP24: Head-Related Transfer Function Dataset Harmonization. *IEEE Open Journal of Signal Processing* 6 (2025), 950–964. doi:10.1109/OJSP.2025.3592601
- [10] Roger R. Davidson. 1970. On Extending the Bradley–Terry Model to Accommodate Ties in Paired Comparison Experiments. *J. Amer. Statist. Assoc.* 65, 329 (1970), 317–328. doi:10.1080/01621459.1970.10481082
- [11] Davide Fantini, Federico Avanzini, Stavros Ntalampiras, and Giorgio Presti. 2021. HRTF Individualization Based on Anthropometric Measurements Extracted from 3D Head Meshes. In *2021 Immersive and 3D Audio: from Architecture to Automotive (3DA)*. 1–10. doi:10.1109/3DA48870.2021.9610904
- [12] Robert H. Gilkey and Timothy R. Anderson (Eds.). 1997. *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum Associates.
- [13] Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [14] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian Active Learning for Classification and Preference Learning. In *arXiv / technical report*. arXiv:1112.5745.
- [15] Darren Irvine, Cal Armstrong, et al. 2019. SADIE II: A New Large HRTF Dataset. In *Proceedings of the AES Conference on Audio for Virtual and Augmented Reality*.
- [16] Brian F. G. Katz and Gaëtan Parseihian. 2012. Perceptually based head-related transfer function database optimization. *The Journal of the Acoustical Society of America* 131, 2 (2012), EL99–EL105. doi:10.1121/1.3672641
- [17] Byeong-Yun Ko, Gyeong-Tae Lee, Hyeonuk Nam, and Yong-Hwa Park. 2023. PRTFNet: HRTF Individualization for Accurate Spectral Cues Using a Compact PRTF. *IEEE Access* 11 (2023), 96119–96130. doi:10.1109/ACCESS.2023.3308143
- [18] Vincent Martin and Lorenzo Picinali. 2025. Impact of HRTF individualisation and head movements in a real/virtual localisation task. arXiv:2510.09161 [eess.AS] https://arxiv.org/abs/2510.09161
- [19] Lucas Maystre and Matthias Grossglauser. 2017. Just sort it! a simple and effective approach to active preference learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (Sydney, NSW, Australia) (ICML '17)*. JMLR.org, 2344–2353.
- [20] Kyla McMullen, Agnieszka Roginska, and Gregory Wakefield. 2012. Subjective Selection of Head-Related Transfer Functions (HRTFs) based on Spectral Coloration and Interaural Time Differences (ITD) Cues. *Journal of the Audio Engineering Society: Audio Engineering Society* 2 (10 2012).
- [21] John C. Middlebrooks. 1999. Individual Differences in External-Ear Transfer Functions Reduced by Scaling in Frequency. *The Journal of the Acoustical Society of America* 106, 3 (1999), 1480–1492.
- [22] John C. Middlebrooks. 1999. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *The Journal of the Acoustical Society of America* 106, 3 (1999), 1493–1510. doi:10.1121/1.427147
- [23] David Poirier-Quinot, Martin S. Lawless, Peter Stitt, and Brian F.G. Katz. 2022. HRTF Performance Evaluation: Methodology and Metrics for Localisation Accuracy and Learning Assessment. In *Advances in Fundamental and Applied Research on Spatial Audio*, Brian F.G. Katz and Piotr Majdak (Eds.). IntechOpen, London, Chapter 3. doi:10.5772/intechopen.104931
- [24] P. V. Rao and L. L. Kupper. 1967. Ties in Paired-Comparison Experiments: A Generalization of the Bradley–Terry Model. *J. Amer. Statist. Assoc.* 62, 317 (1967), 194–204.
- [25] Telecommunications & Signal Processing Laboratory, Multimedia Signal Processing, McGill University. [n. d.]. Speech Database. https://www.mmsp.ece.mcgill.ca/Documents/Data/. Accessed 2025-12-30.
- [26] L. L. Thurstone. 1927. A Law of Comparative Judgment. *Psychological Review* 34, 4 (1927), 273–286. doi:10.1037/h0070288

- [27] Edgar A. Torres-Gallegos, Felipe Orduña-Bustamante, and Fernando Arámbula-Cosío. 2015. Personalization of head-related transfer functions (HRTF) based on automatic photo-anthropometry and inference from a database. *Applied Acoustics* 97 (2015), 84–95. doi:10.1016/j.apacoust.2015.04.009
- [28] Hans Wallach. 1940. The Role of Head Movements and Vestibular and Visual Cues in Sound Localization. *Journal of Experimental Psychology* 27, 4 (1940), 339–368. doi:10.1037/h0054629
- [29] Yuxiang Wang, You Zhang, Zhiyao Duan, and Mark Bocko. 2021. Global HRTF Personalization Using Anthropometric Measures. *Journal of the Audio Engineering Society* 10502 (may 2021).
- [30] Elizabeth M. Wenzel, Marianne Arruda, Dennis J. Kistler, and Frederick L. Wightman. 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 94, 1 (1993), 111–123. doi:10.1121/1.407089
- [31] Dingding Yao, Jiale Zhao, Longbiao Cheng, et al. 2022. An individualization approach for head-related transfer function in arbitrary directions based on deep learning. *JASA Express Letters* 2, 6 (2022), 064401. doi:10.1121/10.0011575
- [32] Rui Zhang, Ruijie Meng, Jinqiu Sang, Yi Hu, Xiaodong Li, and Chengshi Zheng. 2023. Modelling individual head-related transfer function (HRTF) based on anthropometric parameters and generic HRTF amplitudes. *CAAI Transactions on Intelligence Technology* 8, 2 (2023), 364–378. arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/cit2.12196 doi:10.1049/cit2.12196

A Supplementary Materials

The calibration phase (pairwise preference) and evaluation phase (pointing via an azimuth dial and elevation bar, Fig. 3) used different modalities, which reduces direct re-test effects but leaves broader generalization questions open.

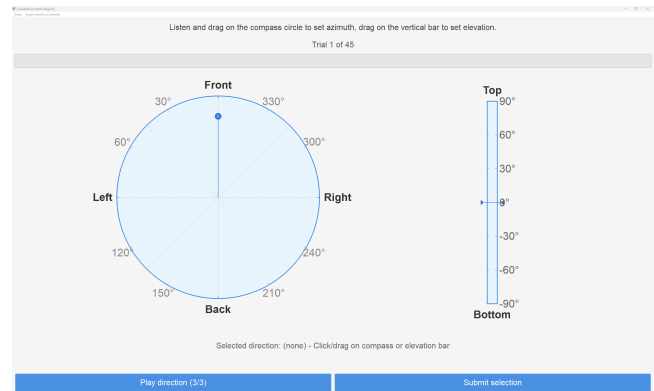


Figure 3: Localization test UI.

Full Bayesian analysis specifications (finite Gaussian mixtures with Dirichlet priors on condition weights, multinomial models for error categories, posterior predictive checks, and Gelman–Rubin convergence diagnostics), supplementary visualizations (response heatmaps, head-pose trajectory clustering, HRTF PCA maps, and anthropometric embeddings), head-motion cluster analyses, and prior specifications with their impact on outcomes are provided in our public repository (Footnote 1).