Comparison of direct and indirect perceptual head-related transfer function selection methods

Franck Zagala, Markus Noisternig, Brian F. G. Katz

JASA ARTICLE

Comparison of direct and indirect perceptual head-related transfer function selection methods

Franck Zagala,[1,a)] Markus Noisternig,[2,b)] and Brian F. G. Katz[1,c)]
[1]Sorbonne Université, CNRS, Institut Jean Le Rond d'Alembert, UMR 7190, Paris, F-75005, France
[2]Sciences et Technologies de la Musique et du Son (STMS) - IRCAM, CNRS, Sorbonne Université, 75004 Paris, France

ABSTRACT:
When a personalized set of head-related transfer functions (HRTFs) is not available, a common solution is identifying a perceptually appropriate substitute from a database. There are various approaches to this selection process whether based on localization cues, subjective evaluations, or anthropomorphic similarities. This study investigates whether HRTF rankings that stem from different selection methods yield comparable results. A perceptual study was carried out using a basic source localization method and a subjective quality judgment method for a common set of eight HRTFs. HRTF rankings were determined according to different metrics from each method for each subject and the respective results were compared. Results indicate a significant and positive mean correlation between certain metrics. The best HRTFs selected according to one method had significant above-average rating scores according to metrics in the second method.
https://doi.org/10.1121/10.0001183
(Received 5 November 2019; revised 3 April 2020; accepted 14 April 2020; published online 11 May 2020)
[Editor: Jonas Braasch] Pages: 3376–3389

NOMENCLATURE

G_x	Individual grading list of HRTFs according to metric x
g_x	Individual grading of HRTFs according to metric x
R_x	Individual ranking list of HRTFs according to metric x
θ	Lateral angle
ϕ	Polar angle
γ	Great circle angle
r_{loc}, r_{traj}	Subject repeatability index for the localization (loc) and trajectory (traj) evaluation tasks, respectively
ρ	Pearson linear correlation coefficient
τ	Kendall's rank correlation coefficient
$\mathcal{D}_{G_1 \to G_2}$	Worst downgrading score according to G_2 of the best selected according to G_1
α	Significance level
p	Probability value
$\bar{\Delta}$	Mean difference

I. INTRODUCTION

In the fields of spatial hearing and spatial audio reproduction, headphone-based techniques have been extensively investigated and used (Begault, 1994; Blauert, 1996; Rumsey, 2012). These techniques aim to convert a given incident sound field into binaural signals such that the sound pressure at each eardrum as produced by the headphones is perceptually equivalent to the original circumstances. This equivalence implies the presence of morphologically determined acoustic cues, such as the low-frequency interaural time difference (ITD), high-frequency envelope ITD, and, to a lesser extent, high-frequency interaural level difference (ILD) for lateral source direction localization, the high-frequency ILD for auditory distance perception of near-field sources outside the median plane, and monaural cues that help discriminate direction on a cone of confusion (i.e., constant ITD or ILD contour; see Katz and Nicol, 2019, Fig. 11.1) through spectral indices. All of these cues are contained in the so-called head-related transfer function (HRTF; Blauert, 2013). The HRTF (e.g., a set of transfer functions measured on a spherical grid at a fixed distance) can be separated into a time delay component, principally influenced by the size of the head and position of the ears, and a spectral component, predominantly determined by the shape of the pinnae, as well as general head shadowing (Katz and Noisternig, 2014).

Practically, in the case of large studies or with specific subject pools (Afonso *et al.*, 2005; Katz and Picinali, 2011; Picinali *et al.*, 2014), the use of an individual HRTF is often impossible as the measurement generally requires (among other difficulties) an anechoic room, specific hardware to position the source on a spherical grid, and tracking devices to ensure the subject remains immobile (Carpentier *et al.*, 2014). For these reasons, other strategies have been proposed to obtain "individual" HRTFs without having to measure them acoustically. Individual HRTFs can be obtained using numerical simulations (Greff and Katz, 2007; Katz,

[a)]Also at Sciences et Technologies de la Musique et du Son (STMS) - IRCAM, CNRS, Sorbonne Université, 75004 Paris, France. Electronic mail: franck.zagala@ircam.fr
[b)]ORCID: 0000-0003-1347-9826.
[c)]ORCID: 0000-0001-5118-0943.

2001a,b; Ziegelwanger *et al*., 2015). One can also transform or tune non-individualized HRTFs (Middlebrooks, 1999; Middlebrooks *et al*., 2000) or select a HRTF from a database by preference or performance (Bahu, 2016; Guillon, 2009; Iwaya, 2006).

A. Methods for selection of HRTF

Due to the large number of publicly available HRTF databases (Algazi *et al*., 2001; Bomhardt *et al*., 2016; SOFA, 2019; Warusfel, 2003), using an already measured non-individualized HRTF offers many practical advantages. However, the choice of the best fitting HRTF from these databases is not trivial. A fundamental and as yet unanswered question is "What determines the suitability of a HRTF for a given subject?" (Katz and Nicol, 2019). Are good HRTFs necessarily characterized by precise localization (direct approaches), or should other subjective evaluations be taken into consideration (Simon *et al*., 2016)?

1. Direct perceptual approaches

As HRTF selection methods are often validated by a localization experiment, it is commonly assumed that the optimal HRTF leads to a minimization of localization error (Geronazzo *et al*., 2019, 2014; Iwaya, 2006; Katz and Parseihian, 2012; Seeber and Fastl, 2003; Voong and Oehler, 2019). Therefore, one can try to identify the HRTF that would lead to a minimization of angular errors or confusion rates. However, such methods typically require specific hardware (for the reporting of perceived direction in the case where a direct pointing method is preferred), are rather time consuming (as many source positions should be evaluated in order to assess the suitability of the HRTF over the whole sphere), and, finally, the results may often be difficult to interpret due to the high variance and multimodal distribution of responses (Bahu *et al*., 2016).

2. Indirect quality judgment approaches

Subjective approaches have been proposed and aim to reduce the complexity of the setup and testing times for HRTF evaluations.

Seeber and Fastl (2003) proposed a two-step approach. The first step consisted in a rapid extraction of 5 HRTFs among an initial set of 12 according to their *overall spatial perception* in the frontal region. For the second step, each subject ranked the remaining HRTFs with respect to different criteria based on externalization, stability of elevation and distance, and constant displacement speed. Results showed in a posterior localization task that this selection led to a reduction in variance in localization responses and the occurrence of inside-the-head localizations.

When dealing with large databases, Iwaya (2006) proposed a Swiss-tournament method where the subject selected the best HRTF out of a pair for a series of paired comparisons. More recently, Voong and Oehler (2019) also proposed a Swiss-tournament method where subjects compared competing HRTFs based on *preference*, *externalization* of two different stimuli, and *envelopment*. This method is well established in chess competitions since it avoids forcing all possible pair combinations to compete together, hence reducing the time of the experiment while also avoiding early elimination of serious competitors that could eventually occur due to answer variance.

Roginska *et al*. (2010) and McMullen *et al*. (2012) proposed a method for identifying a HRTF or a reduced set of HRTFs that subjects would choose more often than others according to three different criteria: *externalization*, *elevation discrimination*, and *front/back discrimination*. Results showed that there existed a subset of the HRTF database that was preferred by a significant number of subjects. Furthermore, groups of subjects with preferences for similar subgroups of HRTFs were highlighted. The authors concluded that the proposed selection method "*would provide very good, and certainly better than generic, cues which would lead to an improved spatial auditory image in virtual environments …*" (Roginska *et al*., 2010).

In order to reduce the size of large databases, Katz and Parseihian (2012) proposed a method where subjects rated overall auditory experience on a three-point scale (*bad*/*ok*/*excellent*), where virtual sources were moving along two different trajectories. Through the analysis, the 46 HRTFs of the database were reduced to 7 HRTFs. The subset appeared to be efficient as results showed that subjects using their best rated HRTF out of the selection method scored significantly better than subjects with their worst rated HRTF during a localization task. Andreopoulou and Katz (2016b) followed a similar approach using a nine-point scale in order to create a perceptually relevant space for quantifying similarities of HRTFs and subjects. Results showed that the HRTFs from the reduced database in Katz and Parseihian (2012) appear to be perceptually orthogonal.

However, despite the various methods, there has been considerable lack of cross-comparison and repeatability test validation of these methods among competing methods. The aim of this study is to provide such a comparison between two common techniques for various metrics.

B. Hypotheses

The current study compares an indirect and a direct perceptual method for the selection of the optimal HRTF from a database. The chosen direct method consists of a classic localization task, whereas the chosen indirect method consists in an overall rendering quality evaluation via a trajectory evaluation task employing a virtual source trajectory around the listener (Andreopoulou and Katz, 2016b; Katz and Parseihian, 2012; Stitt *et al*., 2019). A series of method-specific metrics are evaluated, resulting in quantitative quality assessments for each of the two methods. The goal of the study, therefore, is to compare the ranking of a set of HRTFs according to these two methods. Both methods employed the same HRTF database and the same subject pool.

The hypotheses to be tested are as follows:

$\mathcal{H}_1$ Localization performances across HRTFs are correlated to overall quality of experience judgments.

$\mathcal{H}_2$ The best HRTF selected according to perceptual metrics for one given method will exhibit a rating score better than a random selection in the alternate method.
$\mathcal{H}_3$ Some metrics of a given method are better predictors of the other method's metrics than others.
$\mathcal{H}_4$ Subjects that are most repeatable in one task are also most repeatable in the other task.
$\mathcal{H}_5$ Subjects that are most reliable have the most similar HRTF rankings between both methods.

The description of the experiment is presented in Sec. II, the results are then given in Sec. III, and discussed in Sec. IV.

II. METHOD

A. Description of the experiment

1. Stimuli

The HRTFs used in this work stem from the original LISTEN database (Warusfel, 2003), which initially comprised 46 HRTFs. This database was reduced to 7 HRTFs that ensured that at least 1 HRTF suited each of the 45 subjects of the experiment described in Katz and Parseihian (2012). This reduced database was extended here by including the HRTF of a reference dummy head (Neumann KU100, Neumann, Berlin, Germany), which was measured under the same conditions as the LISTEN database. This resulted in a database of eight HRTFs.

The test stimuli used were a sequence of three white Gaussian noise bursts of 40 ms separated by a 30 ms pause. To avoid undesired artefacts, a 2 ms half Hann window for fade-in and fade-out was applied. This signal is referred to as the *noise burst*.

The headphone (Sennheiser HD 600, Sennheiser, Wedemark, Germany) level was calibrated using an artificial ear (B and K 4153, Brüel & Kjær, Nærum, Denmark), preamplifier (Band K 2669 L), signal conditioner (B and K Nexus), and analog to digital converter (National Instruments NI-USB 9162 and NI 9234, National Instrument, Austin, TX). A microphone calibrator (B and K 4231, Brüel & Kjær, Nærum, Denmark) was used to calibrate the artificial ear's sensitivity at 1 kHz before measurements. The excitation signal was generated such that the stationary white Gaussian noise used to create the *noise burst* (prior to windowing) produced a sound pressure level of 80 dB when averaged over all tested HRTFs and positions.

All binaural signals were played back in a static manner using only source directions made available in the database. No spatial interpolation of the HRTF was required, hence avoiding eventual interpolation parameter/methods-dependent bias in the results.

While inappropriate headphone-to-ear-canal equalization can be a cause of externalization error (Durlach *et al.*, 1992) and lack of naturalness in binaural rendering, such equalization has not been shown to have a significant impact on (angular) localization accuracy (Engel *et al.*, 2019; Schönstein *et al.*, 2008). Any headphone equalization would be universally applied to all positions and is equivalent to a source filter as it is independent of the HRTF difference effects under study.

ITDs of all HRTFs were individualized to each subject using a morphological ITD model based on their measured head circumference, the ITD of each HRTF was modified accordingly (Aussal *et al.*, 2012).

2. Rendering hardware

Binaural signals were generated using the *Anaglyph* audio Plug-in (Poirier-Quinot and Katz, 2018) and played back on open headphones (Sennheiser HD 600, Sennheiser, Wedemark, Germany) using an audio interface (RME Babyface Pro, RME, Heimhausen, Germany) at a sampling rate of 44.1 kHz.

3. Subjects

Twenty-eight subjects took part in the study on a voluntary basis. Subjects' ages ranged from 19 to 63 years old ($\bar{x} = 27.7$, $\sigma^2 = 9.3$) and 10 of the 28 were female. Following a questionnaire, five subjects were identified as *expert listeners* and six others as having experience with binaural audio content.

4. Course of the experiment

The experiment was divided in a series of steps:

- Audiogram in order to identify potential heavy hearing losses of any subject,
- measurement of the head circumference in order to personalize the ITD of each HRTF,
- questionnaire collecting personal information for statistical analysis,
- localization task as described in Sec. II B,
- trajectory evaluation task as described in Sec. II C,
- collection of a series of photographs to extract morphological measures to be used in a future study, and
- short post-task interview asking for eventual remarks about their impressions of the undertaken tasks.

The presentation order between the localization and trajectory evaluation tasks was equally balanced and randomly assigned.

B. Localization task

Subjects indicated the perceived direction of the binaurally rendered noise burst with a direct pointing method without taking distance into account. The experiment took place in virtual reality where subjects were equipped with a head mounted device (HMD; Oculus Rift, Oculus VR, Menlo Park, CA) and two hand controllers (Oculus Touch) that were tracked using three infrared sensors (Oculus Sensor).

The main part of the program was written in C# within the *Unity* framework, which communicated over *open sound control* (Wright, 2005) with Max-MSP (Cycling '74, San Francisco, CA).

Each HRTF was evaluated at 13 different positions, resulting in 8 HRTFs $\times$ 13 positions $=$ 104 different

https://doi.org/10.1121/10.0001183

conditions played in a random sequence within a block. Repetitions were included by concatenating three blocks and ensuring that the first stimulus of each block was not identical to the last of the previous block. The 13 selected source positions, as well as the measurement grid of the *LISTEN* database, are depicted in Fig. 1.

The virtual environment was intentionally left black in order to avoid eventual visual bias on the perceived direction of the stimuli. Furthermore, no avatar was presented such that the subject could not see their hands. That way, any eventual graphic bias between the actual and the visualized hand position due to poor positioning of the HMD was avoided.

Each localization response was decomposed in the following steps:

(1) Subject aligns a small dot placed in front of their head on a distant target (2.2 m radius disk at 100 m = 4.4° aperture angle).
(2) After 1 s of continuous visual/head alignment, the stimulus was triggered, ensuring the subject's head was stable during playback of the short noise burst. This is important as the binaural signal should have a static position. Simultaneously, the position and orientation of the head was logged as the referential for the response. The visual dot and target then disappeared.
(3) The subject could use either the left or right controller to point toward the perceived direction such that the direction coincided with an invisible line starting from the center of the head position[1] from step (2) and an indicated point on the controller. Thus, subjects used their proprioception to place the controller at the response position. The controller reference point was chosen in the center of the small surface between the joystick and two buttons such that subjects could easily place the tip of their thumb on it without needing to see it.
(4) The subject validated the perceived direction by pressing a trigger on the controller with their index finger on the corresponding controller. The controller position relative to the head position in step (2) was logged.

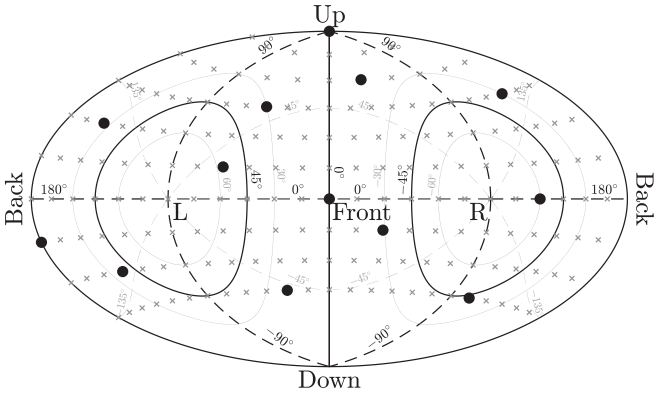

FIG. 1. Hammer projection of the spherical HRTF measurement grid. Solid lines depict constant lateral angles and dashed lines depict constant polar angles. (Black circles) Selected source positions × available source positions.

(5) The visual alignment dot and target reappear, and steps (1)–(5) are repeated until all conditions and repetitions have be completed.

In order to reduce fatigue and provide information to the subjects about their advancement in the session, a short text proposing to take a break appeared between each block. Subjects could either remove the headphones and HMD and go for a short break or alternatively press a button to continue the task after a short 10 s pause.

C. Trajectory evaluation task

The global quality judgment test was similar to that of Stitt *et al.* (2019), wherein subjects rate the set of eight HRTFs on a nine-point discrete scale (the extrema being "worst" and "best") for two different virtual source trajectories, namely a *horizontal plane trajectory* and a *median plane trajectory*.

The graphical user interface (GUI) was implemented in MATLAB (MathWorks, Natick, MA) and presented on a touch screen. The pre-rendered binaural signal was played directly from MATLAB.

For each of the two trajectory types, a short text description of the intended source trajectory was displayed on the left-hand side of the GUI (allowing subjects to form a mental reference). On the right-hand side, the eight stimuli were presented in a random order in the form of a "play/pause" button associated with a set of nine radio buttons for the ratings. Equal ratings between stimuli were accepted and subjects were provided a "sort" button, which reordered the stimuli with respect to the assigned rating at any time during the task.

After subjects listened and rated each stimulus and both extrema were assigned at least once, the "validate" button was made available.

For both trajectories, the stimuli comprised a sequence of noise bursts (described in Sec. II) sequentially rendered at positions directly available in the HRTF (no interpolation) corresponding to the following trajectories:

(a) *Horizontal plane trajectory*. Starting from the front on the horizontal plane and progressing iteratively to the next position 30° counterclockwise, completing two revolutions.
(b) *Median plane trajectory*. Starting at an elevation of −45° on the median plane and progressing iteratively to the next position 15° above until reaching the back of the subject at an elevation of −45°. It then returned to the starting point by completing the trajectory in reverse order.

III. RESULTS

Results for two subjects had to be removed due to a hardware failure that occurred during the localization task, which led to false responses. Both subjects were in the group that began with the trajectory task, had no prior experience with spatial audio or perceptual experiments, and were female.

A. Establishment of HRTFs gradings and rankings

The term *grading* (G) is employed here for normalized values between zero and one that rate each HRTF according

to the metrics described below while preserving the interval relations, while *ranking* (R) is used to regard only the ordinal relation of the grading between HRTFs. The grading of a single HRTF within G (e.g., a selected best HRTF) is represented by the letter g. To highlight eventual correlations between the results of each method, HRTF gradings and rankings were established for each subject with respect to different metrics: *mean unsigned lateral error*, *mean unsigned polar error*, *mean great circle error*, and *confusion rate* for the localization method and *horizontal plane score*, *median plane score*, and *both planes score* for the quality evaluation method. These metrics and the construction of their respective HRTF gradings G_x and rankings R_x are described in Secs. III A 1 and III A 2.

1. Metrics for the localization method

Four metrics are proposed to process the results obtained during the localization task based on different types of localization errors. Each metric produces a grading list for the eight tested HRTFs. For this, the interaural-polar coordinate system (Morimoto and Aokata, 1984) is preferred as depicted in Fig. 2. This coordinate system has the benefit to describe the direction of an object with a lateral angle θ that specifies a cone of confusion (left-hand side at 90°, median plane at 0°, and right-hand side at −90°) and a polar angle ϕ that assesses the position of the object on the cone of confusion (starting at 0° at the front and rotating in an upward direction with increasing ϕ) for a given distance.

a. Mean unsigned lateral error. Computed for each subject for each HRTF by averaging all unsigned lateral angle errors between target position and response over the 3 repetitions and 13 target positions. The corresponding HRTF grading and ranking are referred to as $G_{\bar{\theta}}$ and $R_{\bar{\theta}}$ respectively.

b. Mean unsigned polar error. Computed for each subject for each HRTF by averaging all unsigned polar angle errors between the target position and response over the 3 repetitions and 13 target positions. The corresponding HRTF grading and ranking are referred to as $G_{\bar{\phi}}$ and $R_{\bar{\phi}}$, respectively.

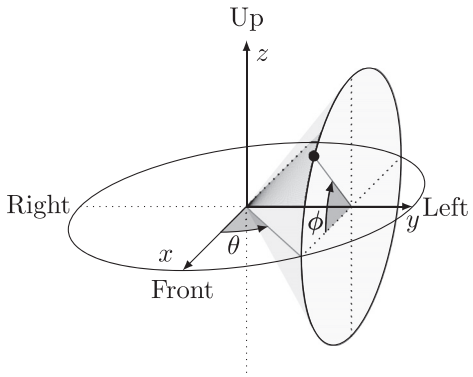

FIG. 2. Interaural-polar coordinate system. A cone of confusion is drawn in light gray and determined by the lateral angle θ. The elevation of the object is given by the polar angle ϕ.

c. Mean great circle error. Computed for each subject for each HRTF by averaging all great circle angle errors between the target position and response over the 3 repetitions and 13 target positions. The corresponding HRTF grading and ranking are referred to as $G_{\bar{\gamma}}$ and $R_{\bar{\gamma}}$, respectively.

d. Confusion rate. Confusion classification follows the definition proposed in Parseihian and Katz (2012), classifying the response directions according to four different error types depending on the polar angle difference between target and response. The classification zones are discussed in Appendix A and shown in Fig. 6 in Appendix A. The different confusions types are grouped, resulting in a single *confusion* region, equal to 1 − precision. The corresponding HRTF grading and ranking are referred to as G_{conf} and R_{conf}, respectively.

HRTF rankings are achieved by sorting the scores in increasing order, thus, the smaller scores are at the *top* of the ranking and the larger scores are at the *bottom* of the ranking. In the case of an equal score, identical ranking positions are given.

In order to give better insight of the absolute accuracy of subjects during the experiment, the confusion rate of each subject, as well as the distribution of absolute errors of each of their responses, is discussed in Appendix B and depicted in Fig. 7 in Appendix B.

2. Metrics for the quality evaluation method

For the trajectory evaluation task, ratings of each HRTF are averaged over the 3 repetitions for each trajectory. In addition to *horizontal plane score* and *median plane score*, a third, *Both planes score*, is the mean result across both trajectories. For consistency with the localization metrics, the resulting gradings of each subject G_{hori}, G_{med}, and G_{both} are determined by norming the abovementioned averaged ratings such that the average *best* rated HRTF receives the grade zero and the *worst* receives the grade one. The resulting HRTF rankings are referred to as R_{hori}, R_{med}, and R_{both} accordingly.

B. Repeatability of the subjects

In order to assess the repeatability of each subject, a *repeatability index* is calculated for each of the two tasks.

r_{loc} As a dispersion indicator, the repeatability index of the localization task is computed by averaging the great circle distance between all three repetitions across all conditions and scaling the result such that the worst possible mean great-circle distance (i.e., 120° for three repetitions) yields zero and perfect agreement yields one.

r_{traj} The repeatability score of each trajectory is computed using Kendall's coefficient of concordance, W (Kendall and Smith, 1939), between the rankings obtained in

https://doi.org/10.1121/10.0001183

each repetition.[2] W ranges from zero to one, where zero indicates strictly no concordance between HRTF rankings and one indicates identical rankings between repetitions. The repeatability indexes of the two trajectories are then averaged to obtain r_{traj}.

The repeatability indices are compared (see Fig. 3) using a linear regression and computing the resulting Pearson linear correlation coefficient ($\rho = 0.51$). The significant positivity of this correlation was tested with a one-tailed t-test and the t-statistics were computed according to Eq. (1) (see Soper *et al.*, 1917).

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}, \tag{1}$$

where n is the sample size. A significant positive linear correlation between both repeatability indices was observed ($\rho = 0.51, p < 0.005$).

The significant positive linear correlation between the repeatability indices of both tasks indicates that subjects who were most consistent during the trajectory task also tended to be the most consistent during the localization task. These results indicate the presence of general "good raters" and "bad raters." Absolute values of the repeatability indices are not directly comparable between tasks, therefore no conclusion could be made about whether one task was more repeatable than the other.

C. Comparison of HRTF metrics

1. Correlation between rankings associated to each metric

a. Correlation. Kendall's τ correlation coefficient (Kendall, 1938) was used as a measure of similarity between two rankings. Kendall's τ was preferred for its ability to take ties into account and to be weighted in order to focus on correlations within the top or bottom of the ranking (Vigna, 2015). The definition of Kendall's τ used in this work is given in Appendix C 1.

For each subject, Kendall's τ is computed to assess the similarity between HRTF rankings depending on different metrics from the two methods. The average τ over all subjects is then computed (Table I and Fig. 8 in Appendix C). A one-tailed t-test ($\alpha = 0.05$) shows that the mean Kendall's τ correlation coefficient over all subjects is significantly positive for the pairs (R_{conf} and R_{hori}), ($R_{\bar{\phi}}$ and R_{hori}), ($R_{\bar{\gamma}}$ and R_{hori}), (R_{conf} and R_{med}), ($R_{\bar{\phi}}$ and R_{med}), ($R_{\bar{\gamma}}$ and R_{med}), (R_{conf} and R_{both}), ($R_{\bar{\phi}}$ and R_{both}), and ($R_{\bar{\gamma}}$ and R_{both}) even though a rapid observation of Fig. 8 shows that negative correlation coefficients regularly appear for some subjects. Results showed no significance for inter-metric correlation coefficients involving $R_{\bar{\theta}}$.

These results indicate a relation between the localization accuracy of HRTFs and the overall quality of experience in the context of sources moving along a given trajectory, except for the *mean unsigned lateral error*. Regarding the other metrics, even if resulting inter-metric correlation coefficients seem relatively low, results are encouraging as it speaks for the fact that the better the scores of a HRTF during the quality evaluation method, the lower the confusion rates, polar errors, and great-circle errors, and inversely the same is true.

b. Top-end/bottom-end correlation. Section III C 1 showed a significant positive correlation between rankings from both methods, however, it is not clear whether this correlation equally spans along the whole ranking, e.g., if bad-scoring HRTFs could be more discernible than well-scoring HRTFs. To investigate whether the correlation of ranking lists differs depending on the region of ranking (see Andreopoulou and Katz, 2016a), one can weight Kendall's τ depending on the rankings itself (Vigna, 2015). The definition of the *top-weighted* and *bottom-weighted* Kendall's τ, as well as the statistical analysis, are given in Appendix C 2.

No significant difference was found between the correlation at the bottom-end and at the top-end of the ranking, except for the pairs (R_{hori} and R_{conf}) and (R_{med} and R_{conf}), suggesting that well-scoring HRTFs according to *confusion rate* have more similar scores, and their differentiation is therefore more subject to response noise than the worse-scoring HRTFs.

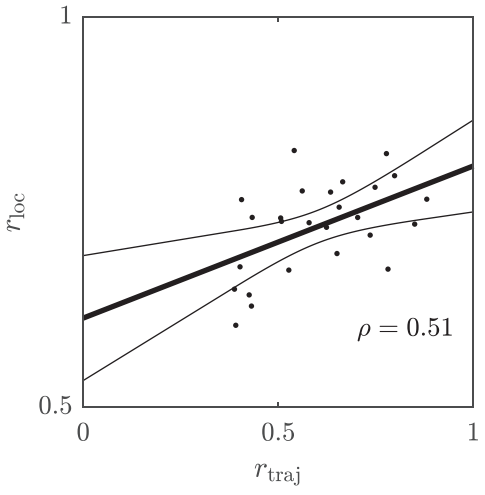

FIG. 3. Repeatability index from the localization task against the repeatability index from the trajectory evaluation task. The linear regression is given with its 95% confidence interval as well as the corresponding Pearson linear correlation coefficient ρ.

TABLE I. Average values of Kendall's τ correlation coefficients between each pair of rankings. "*" indicates a significant positive mean Kendall's τ.

	R_{hori}		R_{med}		R_{both}	
	$\bar{\tau}$	p	$\bar{\tau}$	p	$\bar{\tau}$	p
R_{conf}	0.24*	<0.005	0.16*	0.03	0.23*	<0.005
$R_{\bar{\theta}}$	0.05	0.19	0.10	0.05	0.06	0.12
$R_{\bar{\phi}}$	0.26*	<0.005	0.25*	<0.005	0.30*	<0.005
$R_{\bar{\gamma}}$	0.31*	<0.005	0.27*	<0.005	0.35*	<0.005

c. Comparison of inter-metric correlation coefficients. In order to examine if one metric of the localization method correlates better than the others to a given metric of the quality evaluation method and inversely, correlation coefficients are compared between pairs sharing a common metric using a two-tailed paired t-test. Results indicate that the correlations for the pairs (R_{conf} and R_{hori}), ($R_{\bar{\phi}}$ and R_{hori}), and ($R_{\bar{\gamma}}$ and R_{hori}) are all three significantly greater than ($R_{\bar{\theta}}$ and R_{hori}) ($\bar{\tau} = 0.19$, $p = 0.05$, $\bar{\tau} = 0.22$, $p = 0.01$ and $\bar{\tau} = 0.26$, $p < 0.005$, respectively).

Similarly, it appears that the correlations for the pairs ($R_{\bar{\phi}}$ and R_{med}) and ($R_{\bar{\gamma}}$ and R_{med}) are both significantly greater than ($R_{\bar{\theta}}$ and R_{med}) ($\bar{\Delta} = 0.15$, $p = 0.04$ and $\bar{\Delta} = 0.17$, $p = 0.02$, respectively).

When comparing correlation between metrics from the localization method with *both planes score*, it appears that there is a significantly higher correlation for the pair ($R_{\bar{\phi}}$ and R_{both}) than for ($R_{\bar{\theta}}$ and R_{both}) ($\bar{\Delta} = 0.23$, $p < 0.005$), and there is a higher correlation for the pair ($R_{\bar{\gamma}}$ and R_{both}) than for (R_{conf} and R_{both}) or ($R_{\bar{\theta}}$ and R_{both}) or ($R_{\bar{\phi}}$ and R_{both}) ($\bar{\Delta} = 0.12$, $p = 0.04$; $\bar{\Delta} = 0.29$, $p < 0.005$; and $\bar{\Delta} = 0.12$, $p = 0.02$, respectively).

No significant difference was observed when comparing inter-metric correlation coefficients sharing a common metric from the localization method. However, the correlation for the pair (R_{both} and $R_{\bar{\gamma}}$) is nearly significantly greater than the correlation for the pair (R_{med} and $R_{\bar{\gamma}}$) ($\bar{\Delta} = 0.08$, $p = 0.06$).

This at least suggests that the *mean unsigned polar error* and the *mean great circle error* might be better indicators of the overall experience of a HRTF than the *confusion rate* or the *mean unsigned lateral error*.

On the other hand, *horizontal plane score* and *median plane score* correlate equally well with rankings elaborated from the localization method. However, combining the results from both trajectories increases the correlation with rankings obtained from *mean unsigned polar error* and *mean great circle error* compared to either single trajectory. This indicates that HRTF selection should not be restrained to only one of these trajectories as they seem to deliver complementary information.

d. Relations between inter-metric correlations and repeatability indices. The previously noted negative Kendall's τ (see Sec. III C 1 a) that occurred for some people between rankings associated to different metrics deserves more investigation. To examine whether low inter-metric correlation coefficients correlated to poor repeatability indices, a linear regression was calculated between Kendall's τ obtained for each pair of metrics and repeatability indices of each subject (see Fig. 4). After testing the null hypothesis that correlation indices $\rho_{r_{\text{loc}}}$ and $\rho_{r_{\text{traj}}}$ were non-positive for each pair of metrics with a one-tailed t-test ($\alpha = 0.05$) and computing the t-statistics according to Eq. (1), it appears that inter-metric Kendall's τ has a significant positive correlation with the repeatability index associated to the localization task for the following pairs of metrics: (R_{conf} and R_{med}; $p < 0.005$), (R_{conf} and R_{both}; $p = 0.02$), ($R_{\bar{\phi}}$ and R_{hori}; $p = 0.03$), ($R_{\bar{\phi}}$ and R_{med}; $p = 0.01$), ($R_{\bar{\phi}}$ and R_{both}; $p < 0.005$), ($R_{\bar{\gamma}}$ and R_{med}; $p = 0.03$), and ($R_{\bar{\gamma}}$ and R_{both}; $p = 0.03$). Similarly, the inter-metric Kendall's τ has a significant positive correlation with the repeatability index associated to the trajectory task for the following pairs of metrics: (R_{conf} and R_{med}; $p < 0.005$), (R_{conf} and R_{both}; $p = 0.01$), ($R_{\bar{\phi}}$ and R_{hori}; $p = 0.01$), ($R_{\bar{\phi}}$ and R_{med}; $p < 0.005$), ($R_{\bar{\phi}}$ and R_{both}; $p < 0.005$), ($R_{\bar{\gamma}}$ and R_{hori}; $p = 0.01$), ($R_{\bar{\gamma}}$ and R_{med}; $p < 0.005$), and ($R_{\bar{\gamma}}$ and R_{both}; $p < 0.005$).

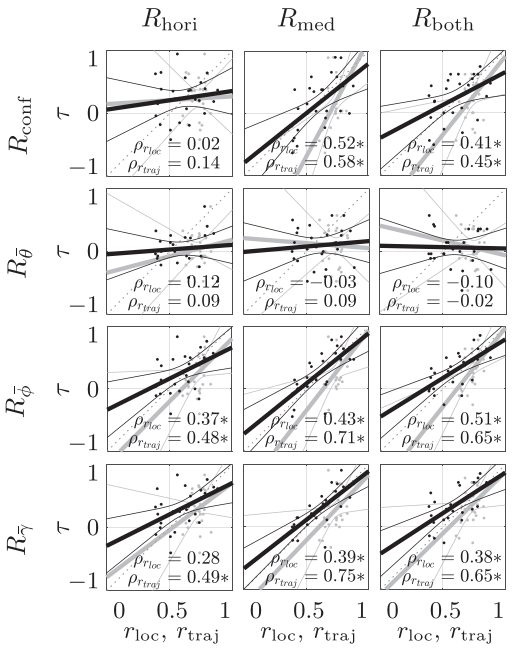

FIG. 4. Kendall's τ against the repeatability score. The Pearson linear correlation coefficient between Kendall's τ and the repeatability score is depicted as well as the linear regression with its 95% confidence interval. (Gray line) Linear regression between Kendall's τ and r_{loc}, (black line) Linear regression between Kendall's τ and r_{traj}. "$*$" indicates a significant positive correlation.

2. Grading behaviour of selected HRTFs

While Sec. III C 1 focused on inter-metric correlation coefficients, observing the behaviour of only the best selected HRTFs across metrics could offer more relevant information in the context of HRTF selection comparisons as a correlation of the whole ranking does not necessarily ensure consistency in the selection results in terms of extrema, e.g., *best*.

To investigate how a selected best HRTF according to one metric is rated according to another metric, a new metric termed the "*worst downgrading score*" ($\mathcal{D}$) is introduced. $\mathcal{D}_{G_1 \to G_2}$ represents the grading score according to G_2 of the

https://doi.org/10.1121/10.0001183

best HRTF selected according to G_1. It is therefore equivalent to $g_2(\text{argbest}\{G_1\})$, where $\text{argbest}\{G_1\}$ denotes the best HRTF according to G_1. A score of $\mathcal{D}_{G_1\to G_2} = 0$ would indicate that the best HRTF in G_1 is also the best HRTF in G_2, while $\mathcal{D}_{G_1\to G_2} = 1$ indicates that the best HRTF in G_1 is the worst HRTF in G_2. The term "worse" takes into account any potential score ties in G_1, in which case only the highest normalized grading in G_2 is retained.

As an example, considering only five elements, if $G_1 = (0, 0.3, 0, 0.4, 1)$ and $G_2 = (0.2, 0.7, 0.5, 0, 1)$, then $\mathcal{D}_{G_1\to G_2} = \max\{0.2, 0.5\} = 0.5$ and $\mathcal{D}_{G_2\to G_1} = 0.4$.

$\mathcal{D}$ is computed for each ordered pair of the gradings from both methods described in Sec. III A. Their distributions are depicted in Fig. 5.

a. Median behaviour of $\mathcal{D}$. In order to determine whether the selection of a best HRTF based on one metric yields a significantly lower median grading according to other metrics compared to random selection (i.e., 0.5), a one-sided Wilcoxon signed rank test ($\alpha = 0.05$) was conducted on all inter-metric $\mathcal{D}$ with results of the significance test shown in Fig. 5.

Analysis of median $\mathcal{D}$ shows that the selection of the best HRTF based on $G_{\bar\phi}$ and $G_{\bar\gamma}$ both lead to a g_{hori} and g_{both} significantly less than 0.5, indicating strong similarity in best HRTF selections:

$$\begin{aligned}
&\text{Med}\{\mathcal{D}_{G_{\bar\phi}\to G_{\text{hori}}}\} = 0.12,\ p < 0.005,\\
&\text{Med}\{\mathcal{D}_{G_{\bar\gamma}\to G_{\text{hori}}}\} = 0.14,\ p < 0.005,\\
&\text{Med}\{\mathcal{D}_{G_{\bar\gamma}\to G_{\text{med}}}\} = 0.16,\ p < 0.005,\\
&\text{Med}\{\mathcal{D}_{G_{\bar\phi}\to G_{\text{both}}}\} = 0.11,\ p = 0.01,\\
&\text{Med}\{\mathcal{D}_{G_{\bar\gamma}\to G_{\text{both}}}\} = 0.06,\ p < 0.005.
\end{aligned}$$

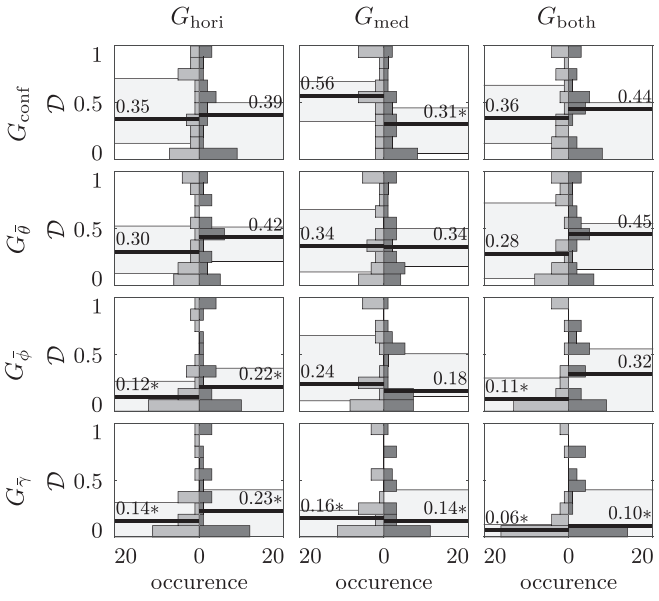

FIG. 5. Histogram representing the distribution of the worst downgrading scores for both directions of each inter-metric pair across all subjects. (Medium gray area) $\mathcal{D}_{G_{\text{row}}\to G_{\text{column}}}$, (dark gray area) $\mathcal{D}_{G_{\text{column}}\to G_{\text{row}}}$, (black line) median value, and (light gray area) 95% confidence interval for the median. "*" indicates a median value significantly smaller than 0.5 ($\alpha = 0.05$).

Therefore, the best HRTF selected based on $G_{\bar\phi}$ and $G_{\bar\gamma}$ would be expected to provide a superior experience in the quality evaluation method compared to a randomly selected HRTF.

Selections of the best HRTF based on G_{hori}, G_{med}, and G_{both} yielded a median $g_{\bar\phi} < 0.5$ and $g_{\bar\gamma} < 0.5$:

$$\begin{aligned}
&\text{Med}(\mathcal{D}_{R_{\text{hori}}\to G_{\bar\phi}}) = 0.22,\ p = 0.02,\\
&\text{Med}(\mathcal{D}_{R_{\text{med}}\to G_{\bar\phi}}) = 0.18,\ p < 0.005,\\
&\text{Med}(\mathcal{D}_{R_{\text{both}}\to G_{\bar\phi}}) = 0.32,\ p < 0.005,\\
&\text{Med}\{\mathcal{D}_{R_{\text{hori}}\to G_{\bar\gamma}}\} = 0.23,\ p = 0.01,\\
&\text{Med}(\mathcal{D}_{R_{\text{med}}\to G_{\bar\gamma}}) = 0.14,\ p < 0.005,\\
&\text{Med}(\mathcal{D}_{R_{\text{both}}\to G_{\bar\gamma}}) = 0.10,\ p < 0.005,
\end{aligned}$$

and seemed to be strong indicators of whether the selected HRTF will score well according to the *mean great circle error*.

Selection based on G_{med} also yielded a median $g_{\text{conf}} < 0.5$ ($\text{Med}\{\mathcal{D}_{R_{\text{med}}\to G_{\text{conf}}}\} = 0.31, p = 0.01$).

These results show that the best HRTF selected based on the *mean unsigned polar error* and *mean great circle error* with the localization method and based on both the *horizontal plane score* and *median plane score* with the quality evaluation method exhibit better scores with the alternate method than randomly selected HRTF.

b. Grading comparison depending on the selection metric. In order to determine which metric of a given method should be prioritized for the selection of the best HRTF such that the grading of the selected HRTF is maximized according to metrics of the alternate method, the distribution of gradings obtained for a given task is compared depending on the metric of the other method used for selection.

First, best HRTF selections were made based on metrics from the trajectory evaluation task and the distributions of the scores $\mathcal{D}_{G\to G_{\text{hori}}}$ for all G in $\{G_{\text{conf}}, G_{\bar\theta}, G_{\bar\phi}, G_{\bar\gamma}\}$ were compared pair-wise as were the distributions of $\mathcal{D}_{G\to G_{\text{med}}}$ and $\mathcal{D}_{G\to G_{\text{both}}}$ (see Table II). Two-tailed paired t-tests ($\alpha = 0.05$) showed that the selected best HRTF based on $G_{\bar\phi}$ exhibits a significantly smaller g_{hori} than the selected best HRTF based on G_{conf} and a nearly significant smaller g_{hori} than the selected best HRTF based on $G_{\bar\theta}$. Furthermore, the selected best HRTF based on $G_{\bar\phi}$ exhibits a significant smaller g_{hori} than the selected best HRTFs based on G_{conf} or $G_{\bar\theta}$.

The selected best HRTFs based on $G_{\bar\phi}$ or $G_{\bar\gamma}$ both exhibit significantly smaller g_{med} than selected best HRTF based on G_{conf}. The selected best HRTF based on $G_{\bar\gamma}$ exhibits significantly smaller g_{med} than the selected best HRTF based on $g_{\bar\theta}$ and nearly significant smaller g_{med} than the selected best HRTF based on $G_{\bar\theta}$.

The selected best HRTF based on $G_{\bar\gamma}$ exhibits significantly smaller g_{both} than the selected best HRTFs based on G_{conf}, $G_{\bar\theta}$, or $G_{\bar\phi}$. Furthermore, the selected best HRTF based on $G_{\bar\phi}$ exhibits a nearly significant smaller g_{both} than the selected best HRTF based on G_{conf}.

On the other hand, when comparing whether the best HRTFs selected based on G_{hori} or G_{med} exhibit better scores

TABLE II. Cross-comparison of $\mathcal{D}$ for different selection metrics across subjects. Results are grouped by grading metric to highlight eventual differences depending on the selection metric. The estimated mean of the difference between paired $\mathcal{D}$ along subjects are given with the corresponding p-values (two-tailed paired t-test with $\alpha = 0.05$). "*" indicates a significant different mean ($\alpha = 0.05$). Gray cells indicate the significantly smaller of the compared $\mathcal{D}$ pairs.

	l	m	$\bar{\Delta}$	p
G_{conf}	$\mathcal{D}_{G_{\text{hori}}\to G_{\text{conf}}}$	$\mathcal{D}_{G_{\text{med}}\to G_{\text{conf}}}$	0.03	0.70
	$\mathcal{D}_{G_{\text{hori}}\to G_{\text{conf}}}$	$\mathcal{D}_{G_{\text{both}}\to G_{\text{conf}}}$	−0.01	0.88
	$\mathcal{D}_{G_{\text{med}}\to G_{\text{conf}}}$	$\mathcal{D}_{G_{\text{both}}\to G_{\text{conf}}}$	−0.05	0.26
$G_{\bar{\theta}}$	$\mathcal{D}_{G_{\text{hori}}\to G_{\bar{\theta}}}$	$\mathcal{D}_{G_{\text{med}}\to G_{\bar{\theta}}}$	0.01	0.94
	$\mathcal{D}_{G_{\text{hori}}\to G_{\bar{\theta}}}$	$\mathcal{D}_{G_{\text{both}}\to G_{\bar{\theta}}}$	−0.02	0.80
	$\mathcal{D}_{G_{\text{med}}\to G_{\bar{\theta}}}$	$\mathcal{D}_{G_{\text{both}}\to G_{\bar{\theta}}}$	−0.02	0.70
$G_{\bar{\phi}}$	$\mathcal{D}_{G_{\text{hori}}\to G_{\bar{\phi}}}$	$\mathcal{D}_{G_{\text{med}}\to G_{\bar{\phi}}}$	0.03	0.70
	$\mathcal{D}_{G_{\text{hori}}\to G_{\bar{\phi}}}$	$\mathcal{D}_{G_{\text{both}}\to G_{\bar{\phi}}}$	0.03	0.72
	$\mathcal{D}_{G_{\text{med}}\to G_{\bar{\phi}}}$	$\mathcal{D}_{G_{\text{both}}\to G_{\bar{\phi}}}$	−0.01	0.87
$G_{\bar{\gamma}}$	$\mathcal{D}_{G_{\text{hori}}\to G_{\bar{\gamma}}}$	$\mathcal{D}_{G_{\text{med}}\to G_{\bar{\gamma}}}$	0.04	0.66
	$\mathcal{D}_{G_{\text{hori}}\to G_{\bar{\gamma}}}$	$\mathcal{D}_{G_{\text{both}}\to G_{\bar{\gamma}}}$	0.06	0.44
	$\mathcal{D}_{G_{\text{med}}\to G_{\bar{\gamma}}}$	$\mathcal{D}_{G_{\text{both}}\to G_{\bar{\gamma}}}$	0.02	0.78
G_{hori}	$\mathcal{D}_{G_{\text{conf}}\to G_{\text{hori}}}$	$\mathcal{D}_{G_{\bar{\theta}}\to G_{\text{hori}}}$	0.03	0.75
	$\mathcal{D}_{G_{\text{conf}}\to G_{\text{hori}}}$	$\mathcal{D}_{G_{\bar{\phi}}\to G_{\text{hori}}}$	0.20	0.03*
	$\mathcal{D}_{G_{\text{conf}}\to G_{\text{hori}}}$	$\mathcal{D}_{G_{\bar{\gamma}}\to G_{\text{hori}}}$	0.21	0.01*
	$\mathcal{D}_{G_{\bar{\theta}}\to G_{\text{hori}}}$	$\mathcal{D}_{G_{\bar{\phi}}\to G_{\text{hori}}}$	0.17	0.03*
	$\mathcal{D}_{G_{\bar{\theta}}\to G_{\text{hori}}}$	$\mathcal{D}_{G_{\bar{\gamma}}\to G_{\text{hori}}}$	0.18	0.07
	$\mathcal{D}_{G_{\bar{\phi}}\to G_{\text{hori}}}$	$\mathcal{D}_{G_{\bar{\gamma}}\to G_{\text{hori}}}$	0.01	0.88
G_{med}	$\mathcal{D}_{G_{\text{conf}}\to G_{\text{med}}}$	$\mathcal{D}_{G_{\bar{\theta}}\to G_{\text{med}}}$	0.14	0.11
	$\mathcal{D}_{G_{\text{conf}}\to G_{\text{med}}}$	$\mathcal{D}_{G_{\bar{\phi}}\to G_{\text{med}}}$	0.16	0.03*
	$\mathcal{D}_{G_{\text{conf}}\to G_{\text{med}}}$	$\mathcal{D}_{G_{\bar{\gamma}}\to G_{\text{med}}}$	0.30	0.00*
	$\mathcal{D}_{G_{\bar{\theta}}\to G_{\text{med}}}$	$\mathcal{D}_{G_{\bar{\phi}}\to G_{\text{med}}}$	0.03	0.77
	$\mathcal{D}_{G_{\bar{\theta}}\to G_{\text{med}}}$	$\mathcal{D}_{G_{\bar{\gamma}}\to G_{\text{med}}}$	0.16	0.08
	$\mathcal{D}_{G_{\bar{\phi}}\to G_{\text{med}}}$	$\mathcal{D}_{G_{\bar{\gamma}}\to G_{\text{med}}}$	0.14	0.02*
G_{both}	$\mathcal{D}_{G_{\text{conf}}\to G_{\text{both}}}$	$\mathcal{D}_{G_{\bar{\theta}}\to G_{\text{both}}}$	0.05	0.61
	$\mathcal{D}_{G_{\text{conf}}\to G_{\text{both}}}$	$\mathcal{D}_{G_{\bar{\phi}}\to G_{\text{both}}}$	0.18	0.07
	$\mathcal{D}_{G_{\text{conf}}\to G_{\text{both}}}$	$\mathcal{D}_{G_{\bar{\gamma}}\to G_{\text{both}}}$	0.28	0.00*
	$\mathcal{D}_{G_{\bar{\theta}}\to G_{\text{both}}}$	$\mathcal{D}_{G_{\bar{\phi}}\to G_{\text{both}}}$	0.12	0.13
	$\mathcal{D}_{G_{\bar{\theta}}\to G_{\text{both}}}$	$\mathcal{D}_{G_{\bar{\gamma}}\to G_{\text{both}}}$	0.23	0.02*
	$\mathcal{D}_{G_{\bar{\phi}}\to G_{\text{both}}}$	$\mathcal{D}_{G_{\bar{\gamma}}\to G_{\text{both}}}$	0.10	0.05*

according to the metrics of the localization method, no significant differences were observed.

These results indicate that when selecting the best HRTF via the localization method, the *mean unsigned polar error* and *mean great circle error* should be preferred over the *confusion rate* and *mean unsigned lateral error* as the selected HRTF will, in general, exhibit better scores according to the metrics of the quality evaluation method. This is consistent with the results obtained when comparing the correlation coefficients of the inter-metrics rankings (see Sec. III C 1 c).

Regarding the gradings from the localization method, there was no significant difference between selections based on the *horizontal plane score* and *median plane score*. This is consistent with observations in Sec. III C 1 c regarding the inter-metric correlation coefficients. However, unlike the improvement of the correlation coefficient shown in Sec. III C 1 c, the best HRTF selection based on the *both planes score* did not appear to significantly improve scores in the localization method compared to selections from a single trajectory.

c. Asymmetry of $\mathcal{D}$. Figure 5 shows that $\mathcal{D}$ is asymmetric, which means that $\mathcal{D}_{G_1\to G_2}$ and $\mathcal{D}_{G_2\to G_1}$, in general, do not give the same score. It thus may be of interest to determine whether a given direction of comparison provided more robust scores for a given pair of metrics. Results are shown in Table III. A two-tailed paired t-test indicated that selecting the best HRTF based on G_{med} generally yielded a significantly smaller G_{conf} than conversely ($\bar{\Delta} = -0.22$, $p = 0.01$). Those observations show that the *median plane score* is a better predictor of the *confusion rate* than is the *confusion rate* a better predictor of the *median plane score*.

D. Open remarks from the subjects

1. Localization task

Among the recurring open remarks on the localization task, 14 subjects spontaneously reported difficulties localizing sources in the front. Five subjects reported that the task was either too long or difficult in general, or expressed their difficulties keeping their concentration throughout the entire task. In contrast, four subjects found the task enjoyable or preferred it over the trajectory evaluation task. Four subjects reported having trouble indicating the perceived direction of the source when the signal was poorly externalized.

2. Trajectory evaluation task

Five subjects found the task difficult in general. Some subjects reported that one of the two trajectories was easier to rate than the other, i.e., seven found the horizontal plane trajectory easier to rate, while only one subject indicated the opposite. Four subjects found the stimulus too loud or irritating. Three subjects reported difficulties localizing the source

TABLE III. Comparison of direction for $\mathcal{D}$ for all inter-method-metric pairs across subjects. The estimated mean of the difference between paired $\mathcal{D}$ along subjects IS given with the corresponding p-values (two-tailed paired t-test with $\alpha = 0.05$). "*" indicates a significant different mean ($\alpha = 0.05$). Gray cells indicate the significantly smaller of the compared $\mathcal{D}$ pair.

l	m	$\bar{\Delta}$	p
$\mathcal{D}_{G_{\text{conf}}\to G_{\text{hori}}}$	$\mathcal{D}_{G_{\text{hori}}\to G_{\text{conf}}}$	0.05	0.63
$\mathcal{D}_{G_{\text{conf}}\to G_{\text{med}}}$	$\mathcal{D}_{G_{\text{med}}\to G_{\text{conf}}}$	0.22	0.01*
$\mathcal{D}_{G_{\text{conf}}\to G_{\text{both}}}$	$\mathcal{D}_{G_{\text{both}}\to G_{\text{conf}}}$	0.06	0.54
$\mathcal{D}_{G_{\bar{\theta}}\to G_{\text{hori}}}$	$\mathcal{D}_{G_{\text{hori}}\to G_{\bar{\theta}}}$	−0.01	0.93
$\mathcal{D}_{G_{\bar{\theta}}\to G_{\text{med}}}$	$\mathcal{D}_{G_{\text{med}}\to G_{\bar{\theta}}}$	0.03	0.75
$\mathcal{D}_{G_{\bar{\theta}}\to G_{\text{both}}}$	$\mathcal{D}_{G_{\text{both}}\to G_{\bar{\theta}}}$	−0.03	0.72
$\mathcal{D}_{G_{\bar{\phi}}\to G_{\text{hori}}}$	$\mathcal{D}_{G_{\text{hori}}\to G_{\bar{\phi}}}$	−0.10	0.25
$\mathcal{D}_{G_{\bar{\phi}}\to G_{\text{med}}}$	$\mathcal{D}_{G_{\text{med}}\to G_{\bar{\phi}}}$	0.10	0.13
$\mathcal{D}_{G_{\bar{\phi}}\to G_{\text{both}}}$	$\mathcal{D}_{G_{\text{both}}\to G_{\bar{\phi}}}$	−0.03	0.62
$\mathcal{D}_{G_{\bar{\gamma}}\to G_{\text{hori}}}$	$\mathcal{D}_{G_{\text{hori}}\to G_{\bar{\gamma}}}$	−0.09	0.21
$\mathcal{D}_{G_{\bar{\gamma}}\to G_{\text{med}}}$	$\mathcal{D}_{G_{\text{med}}\to G_{\bar{\gamma}}}$	−0.01	0.91
$\mathcal{D}_{G_{\bar{\gamma}}\to G_{\text{both}}}$	$\mathcal{D}_{G_{\text{both}}\to G_{\bar{\gamma}}}$	−0.09	0.19

https://doi.org/10.1121/10.0001183

in the front. Finally, one subject admitted preferring to accomplish this task over the localization task.

E. Task duration

The time spent on each task was recorded for 22 subjects without taking pauses into account. The average time spent on the localization task was 24 min 42 s (ranging from 12 min 19 s to 43 min 18 s), while the time spent on the trajectory evaluation task was 26 min 53 s (ranging from 16 min 04 s to 49 min 25 s). Results show the two protocols with the associated number of repetitions comparable in task duration.

IV. DISCUSSION

This study investigated the similarity between two different methods of HRTF rating and selection. A common pool of 28 subjects took part in a classical localization task and a trajectory evaluation task to evaluate 8 HRTFs from a database.

Distribution of angular errors obtained from the localization task (see Appendix B and Fig. 7) were in agreement with the literature. Observed *mean unsigned lateral errors* were similar to those obtained with individual HRTFs (Stitt *et al.*, 2019, results prior to training), although *mean unsigned lateral errors* obtained by four subjects with their best scoring HRTFs were definitely worse. This could be explained by a bad matching of the ITD model for those subjects. *Mean unsigned polar errors* obtained with each subject's best scoring HRTF were scoring similarly to individual HRTFs in Stitt *et al.* (2019), however, worst scoring HRTFs were scoring substantially worse than individual HRTFs. Furthermore, for most of the subjects, a great difference between the median errors of best and worst scoring HRTFs was observed for the *mean unsigned polar error* and *mean great circle error*, indicating that the set of HRTFs was suitable to this work.

Two different approaches were used to compare the similitude of the HRTF scores according to different metrics from both methods. The first approach consisted in computing Kendall's correlation coefficient τ of the HRTF rankings. The second approach consisted in selecting HRTFs based on a given metric and analyzing their corresponding gradings according to other metrics for which a new metric $\mathcal{D}$ was proposed to focus on the change in rating of the best selected HRTF using one metric according to another metric. While the first approach gives insight on the overall behaviour of gradings, the second approach gives confirmation on whether this correlation is valid for selecting the best rated HRTF. Both approaches showed similar results: analyzing inter-method metrics correlation coefficients and $\mathcal{D}$ yielded results significantly above random, supporting $\mathcal{H}_1$ and $\mathcal{H}_2$.

Regarding hypothesis $\mathcal{H}_3$, results showed that some metrics from the localization method (i.e., *mean great circle error* and *mean unsigned polar error*) correlated better to metrics from the quality evaluation method than others when analyzing the full rating results. A similar conclusion was drawn when using the $\mathcal{D}$-approach, which focused only on the best rated HRTFs. This indicates that *the mean great circle error* or *mean unsigned polar error* should be preferred over the *confusion rate* or *mean unsigned lateral error* when selecting HRTFs as it appears that the best selected HRTF will exhibit better scores according to metrics from the quality evaluation method. While HRTF rankings obtained from the *both planes score* appear to better correlate with rankings obtained via the localization method than those based on *the horizontal plane score* or *median plane score* (Table I, Fig. 8 in Appendix C), no statistically significant improvement was observed (at the $\alpha = 0.5$ level). Similarly, when using the $\mathcal{D}$-approach, a greater similitude can be visually observed between the *both planes score* and metrics from the localization method than the *horizontal plane score* or *median plane score* (Fig. 5) even though no statistically significant improvement was observed.

These results are in agreement with observations in previous studies (Iwaya, 2006; Katz and Parseihian, 2012; Seeber and Fastl, 2003; Voong and Oehler, 2019) where localization performances with HRTFs selected using a quality evaluation method were superior to randomly selected HRTFs or generic HRTFs.

As expected, the *mean unsigned lateral error* appears completely uncorrelated to results obtained during the quality evaluation method. This was as expected as is explained by the fact that ITDs were matched to each subject for all HRTFs such that HRTFs could not be discriminated in terms of lateral localization.

Regarding hypothesis $\mathcal{H}_4$, the repeatability of subjects was evaluated for both tasks. For the localization task, the selected repeatability index was inversely proportional to the mean great circle angle between repetitions. For the trajectory evaluation task, Kendall's coefficient of concordance W was used. A significant correlation was found between these repeatability indices, indicating that consistent raters in one task tended to be consistent in the other task.

Inter-metric correlations were observed to significantly correlate to repeatability indices, indicating that consistent raters tended to score best with the same HRTFs in both methods, while inconsistent raters could score differently with each HRTF depending on the method, supporting $\mathcal{H}_5$.

V. CONCLUSION

This study compared two different methods of perceptual HRTF rating and selection from a database containing eight perceptually different HRTFs. The first method was based on a classic localization task, while the second method was based on an overall quality evaluation of sources moving along two different trajectories.

Results showed a significant and positive correlation between rankings established from both methods using different metrics (excluding the *mean unsigned lateral error* from the localization method). Certain metrics calculated from the localization method were observed to better correlate to those from the quality evaluation

method (namely the *mean unsigned polar error* and *mean great circle error)*. Similarly, metrics calculated from the quality evaluation method (*horizontal plane score*, *median plane score*) correlated equally well with those from the localization method, suggesting that the trajectory evaluation task could be reduced to a single trajectory. However, a nearly significant improvement was observed when metrics considered the combined trajectory results, suggesting that employing both trajectories was complementary for this purpose.

A new function was proposed to evaluate the similarity of a selected best HRTF using one metric with respect to another metric, offering results similar to those obtained via correlation coefficient results. The best HRTFs selected according to the *horizontal plane score*, *median plane score*, and *both planes score* were significantly better rated according to the *mean great circle error* for the trajectory evaluation task than randomly selected HRTFs.

It can be claimed that the localization performance of a HRTF positively correlates with the overall quality of experience judgment, evaluated here in the context of sources moving in an anechoic environment.

Results also showed that subjects that are the most repeatable in one task were also the most repeatable in the other task, indicating the presence of "consistent raters" and "inconsistent raters."

Finally, open remarks from subjects showed that preferences and perceived difficulties regarding the two tasks were mixed.

Future research shall extend comparisons of HRTF ranking and selection to objective methods employing anthropomorphic data. Additionally, results of the rankings for the dummy head will be analyzed in an effort to quantify the "genericness" of such systems in the context of binaural listening versus using individually selected HRTFs from a reduced database.

ACKNOWLEDGMENTS

This work was funded in part by the RASPUTIN project (Grant No. ANR-18-CE38-0004).[3] Additional support for B.F.G.K. was provided through a fundamental research collaboration partnership between Sorbonne Université, CNRS, Institut d'Alembert and Facebook Reality Labs.

APPENDIX A: CONFUSION ZONES

A revised version of the definition of the four different response confusion zones proposed in Parseihian and Katz (2012) is given in Fig. 6, in which a correction for south-pole continuity has been added.

APPENDIX B: INDIVIDUAL SUBJECT RESPONSES FOR THE LOCALIZATION TASK

In order to provide a more in-depth analysis of individual subject performance for the localization task, the confusion rates and overall angular error results are shown for each subject in Fig. 7. For comparison, these results are shown in reference to the mean and median values for individual HRTFs using a comparable test protocol (see Stitt *et al.*, 2019, results prior to training).[4]

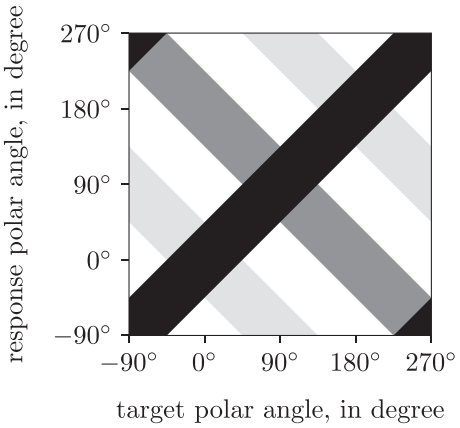

FIG. 6. Definition of the four different response confusion zones. (Black) Precision, (dark gray) front-back, (light gray) up-down, and (white) combined.

Median unsigned lateral error results by subject for their *best* and *worst* scoring HRTFs appear to be generally quite comparable to the mean polar errors observed for the reference individual HRTF study. It can be observed that four subjects appear to have substantially higher errors. This could be due to errors in the ITD individualization model for these specific subjects or generally poor localization task performance.

Regarding median unsigned polar errors, the best scoring HRTFs of most subjects yield median errors comparable to those obtained in Stitt *et al.* (2019). Furthermore, differences obtained between best scoring and worst scoring HRTFs appear to be substantial, indicating that the set of HRTFs used in the current study provides both well matching as well as poorly matching HRTFs for most subjects.

APPENDIX C: KENDALL'S τ

1. Unweighted Kendall's τ

Kendall's τ can be understood as a measure comparing the ordinal relation of each pair of elements, increasing whenever their relation is concordant in both rankings, decreasing in the opposite case, or stable whenever the two elements are tied in one of the two rankings. It is therefore closely related to the minimum number of "swaps" between adjacent elements in a given ranking $\boldsymbol{r}$ in order to reconstruct a ranking $\boldsymbol{s}$.

After considering $\boldsymbol{r}$ and $\boldsymbol{s}$ to be real valued vectors (e.g., the ranking of each HRTF for two given metrics), the original Kendall's τ can be extended to handle ties and is expressed in a form similar to an inner product (Daniels, 1944; Kendall, 1945; Vigna, 2015) as shown:

https://doi.org/10.1121/10.0001183

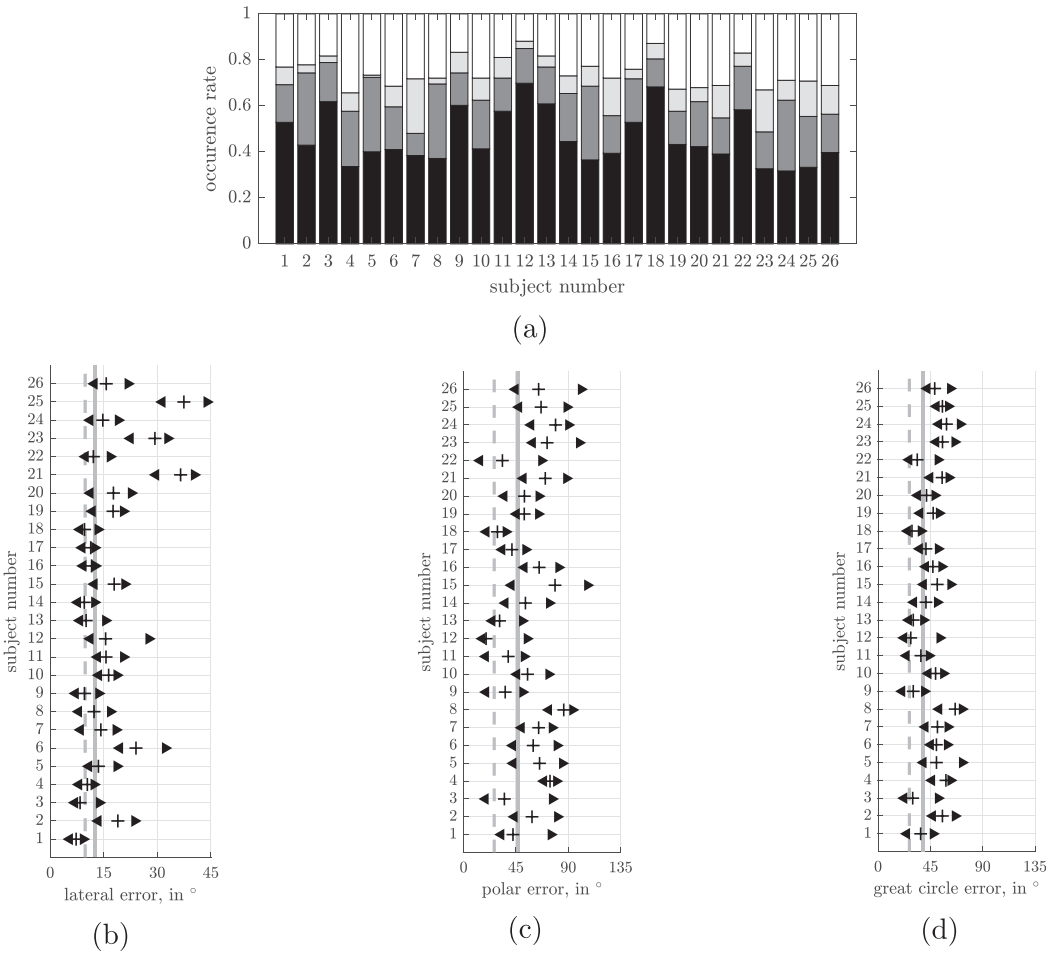

FIG. 7. Distribution of individual subject responses for the localization experiment. (a) Confusion rates. (Black) Precision, (dark gray) front-back, (light gray) up-down, and (white) combined. Median (b) lateral, (c) polar, and (d) great circle errors: (plus sign) over *all* HRTFs, for the (left-pointing triangle) *best* and (right-pointing triangle) *worst* scoring HRTF. (Gray dashed line) Median and (gray solid line) mean errors observed in individual HRTF conditions (from Stitt *et al.*, 2019, results prior to training) are included for reference.

$$\tau = \frac{\langle \boldsymbol{r}, \boldsymbol{s} \rangle}{\|\boldsymbol{r}\| \cdot \|\boldsymbol{s}\|}, \tag{C1}$$

where

$$\langle \boldsymbol{r}, \boldsymbol{s} \rangle = \sum_{i<j} \operatorname{sgn}\{r_i - r_j\} \cdot \operatorname{sgn}\{s_i - s_j\}, \tag{C2}$$

and the normalization is given by the terms

$$\|\boldsymbol{r}\| = \sqrt{\langle \boldsymbol{r}, \boldsymbol{r} \rangle} \quad \text{and} \quad \|\boldsymbol{s}\| = \sqrt{\langle \boldsymbol{s}, \boldsymbol{s} \rangle},$$

and

$$\operatorname{sgn}\{x\} = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x = 0, \\ -1, & \text{if } x < 0. \end{cases}$$

The normalization of the "inner product" in Eq. (C1) ensures that τ reaches one when both rankings are fully concordant, zero whenever there are as many concordant pairs as discordant pairs, and -1 when the two rankings appear in the exact opposite order.

2. Weighted Kendall's τ

Equation (C1) can be re-written by using an alternate definition of Eq. (C2) with

$$\langle \boldsymbol{r}, \boldsymbol{s} \rangle_{w_{r,s}} = \sum_{i<j} \operatorname{sgn}\{r_i - r_j\} \operatorname{sgn}\{s_i - s_j\} w_{\boldsymbol{r},\boldsymbol{s}}(i,j), \tag{C3}$$

where $w_{\boldsymbol{r},\boldsymbol{s}}(i,j)$ is a weighting function which depends on the ranking of the ith and jth elements in $\boldsymbol{r}$ and untied with the help of $\boldsymbol{s}$ if necessary. The weighting of both the ith and jth elements can be combined additively such that $w_{\boldsymbol{r},\boldsymbol{s}}(i,j) = w_{\boldsymbol{r},\boldsymbol{s}}(i) + w_{\boldsymbol{r},\boldsymbol{s}}(j)$.

Thus, one can define two weighted Kendall's τ's, where the weighting function is either based first on the ranking scheme $\boldsymbol{r}$ and then on $\boldsymbol{s}$ ($\tau_{w_{r,s}}$) or inversely ($\tau_{w_{s,r}}$). The final definition of the weighted Kendall's τ takes the form

$$\tau_w = \frac{\tau_{w_{r,s}} + \tau_{w_{s,r}}}{2}. \tag{C4}$$

The choice of the weighting function w remains an open question. Vigna (2015) proposed using a hyperbolic decreasing function $[w : i \to 1/(R(i)+1)]$. However, this

may not be suitable since it introduces a bias, i.e., the mean weighted Kendall's τ is positive even for randomly generated data. Hence, we propose using two very simplistic and intuitive weighting functions, i.e., a *top-weighting* function, where the best rated four HRTFs are weighted one and the four last are weighted zero and a *bottom-weighting* function which is similarly defined but puts the weight onto the four worst rated HRTF, resulting in a *top-weighted Kendall's* τ and a *bottom-weighted Kendall's* τ, respectively.

As an example, consider two rankings $\boldsymbol{r}$ and $\boldsymbol{s}$, where $\boldsymbol{s}$ is similar to $\boldsymbol{r}$ but with two elements swapped; if the two swapped elements lie within the four worst rated elements, then the top-weighted Kendall's τ remains one, however, if at least one of the two swapped elements lie within the four best rated elements, then the top-weighted Kendall's τ decreases with an increasing rankings distance. Thus, if only two elements are swapped, the worst top-weighted Kendall's τ is obtained when swapping the best HRTF with the worst one.

a. Top-end/bottom-end correlation comparison

The top- and bottom-weighted Kendall's τ's between the rankings that stem from different metrics are shown in Fig. 8. The null hypothesis that the correlation at the bottom-end of the ranking is equal to the correlation at the top-end for each pair of metrics was tested using a two-tailed paired t-test ($\alpha = 0.05$). Significance was found only for the pairs (R_{hori} and R_{conf}; $\bar{\Delta} = -0.10$, $p = 0.03$), and (R_{med} and R_{conf}; $\bar{\Delta} = -0.10$, $p = 0.04$), where the correlation at the bottom-end appears greater than at the top-end. These results indicate that the positive mean Kendall's τ observed before might be better explained by the bottom-end of the ranking scale than by the top-end for those cases. Interestingly, both pairs are related to R_{conf}. Therefore, it can be hypothesized that the bottom-end ranking that stems from the *confusion rate* is more stable than at the top-end, i.e., well-scoring HRTFs according to the *confusion rate* have more similar scores, and their differentiation is therefore more subject to response noise than the worse-scoring HRTFs. In all other cases, the positive inter-metric correlation seems as equally explained by the top of the ranking as by the bottom of the ranking.

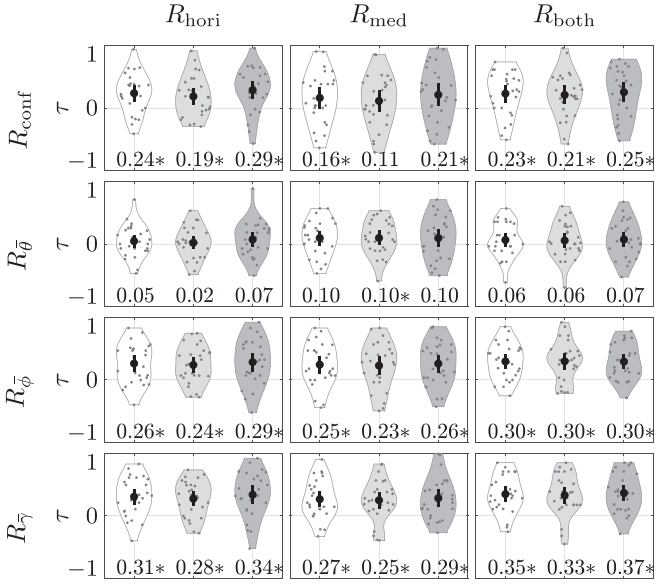

FIG. 8. Kernel density estimation, 95% confidence interval, and average value depicting the distribution of different correlation coefficients between each pair of rankings. The value at the bottom of each violin plot consists of the average correlation value along all subjects. (white) [Kendall's τ, (light gray) top-weighted Kendall's τ, (dark gray) bottom-weighted Kendall's τ. "*" indicates a significant positive mean Kendall's τ.

[1]The center of the head position was determined by averaging the recorded positions of the two hand controllers when placing the controller reference point in front of the ear canal of both ears for 21 subjects.

[2]Kendall's coefficient of concordance, W, is equivalent to a scaled Friedman statistics $W = \text{Fr}/(m(n-1))$, where m is the number of repetitions and n is the number of HRTFs, resulting in a value ranging between zero and one.

[3]See https://rasputin.lam.jussieu.fr (Last viewed 30 April 2020).

[4]The median lateral error and the mean and median great circle errors were not available in Stitt *et al.* (2019) and were therefore computed based on the original data set delivered by the authors.

Afonso, A., Katz, B. F. G., Blum, A., Jacquemin, C., and Denis, M. (**2005**). "A study of spatial cognition in an immersive virtual audio environment: Comparing blind and blindfolded individuals," in *Int. Conf. Auditory Display*, pp. 1–8.

Algazi, V. R., Duda, R. O., Thompson, D. W., and Avendaño, C. (**2001**). "The CIPIC HRTF database," in *IEEE Workshop on the Appl. Sig. Proc. to Audio and Acoust.*, pp. 99–102.

Andreopoulou, A., and Katz, B. F. G. (**2016a**). "Investigation on subjective HRTF rating repeatability," in *Audio Eng. Soc. Conv.*, Vol. 140, pp. 1–10.

Andreopoulou, A., and Katz, B. F. G. (**2016b**). "Subjective HRTF evaluations for obtaining global similarity metrics of assessors and assessees," J. Multimodal User Interfaces **10**(3), 259–271.

Aussal, M., Alouges, F., and Katz, B. F. G. (**2012**). "ITD interpolation and personalization for binaural synthesis using spherical harmonics," in *Aud. Eng. Soc. UK Conf.*, pp. 4.01–4.10.

Bahu, H. (**2016**). "Localisation auditive en contexte de synthèse binaurale non-individuelle" ("Sound localization in the context of non-individualized binaural synthesis"), Ph.D. thesis, Université Pierre et Marie Curie–Paris VI, Paris, France.

Bahu, H., Carpentier, T., Noisternig, M., and Warusfel, O. (**2016**). "Comparison of different egocentric pointing methods for 3D sound localization experiments," Acta Acust. Acust. **102**(1), 107–118.

Begault, D. R. (**1994**). *3-D Sound for Virtual Reality and Multimedia* (Academic, Cambridge).

Blauert, J. (**1996**). *Spatial Hearing* (MIT Press, Cambridge), p. 512.

Blauert, J., ed. (**2013**). "Modern acoustics and signal processing," in *The Technology of Binaural Listening* (Springer, Berlin), p. 511.

Bomhardt, R., de la Fuente Klein, M., and Fels, J. (**2016**). "A high-resolution head-related transfer function and three-dimensional ear model database," Proc. Meet. Acoust. **29**, 050002.

Carpentier, T., Bahu, H., Noisternig, M., and Warusfel, O. (**2014**). "Measurement of a head-related transfer function database with high spatial resolution," in *Forum Acusticum*, [European Acoustics Association (EAA), Krakow, Poland], pp. 1–6.

Daniels, H. E. (**1944**). "The relation between measures of correlation in the universe of sample permutations," Biometrika **33**(2), 129–135.

Durlach, N. I., Rigopulos, A., and Pang, X. D. (**1992**). "On the externalization of auditory image," Presence **1**(2), 251–257.

Engel, I., Alon, D. L., Robinson, P. W., and Mehra, R. (**2019**). "The effect of generic headphone compensation on binaural rendering," in *Aud. Eng. Soc. UK Conf. Imm. Interact. Aud.*, p. 10.

Geronazzo, M., Peruch, E., Prandoni, F., and Avanzini, F. (**2019**). "Applying a single-notch metric to image-guided head-related transfer function selection for improved vertical localization," J. Audio Eng. Soc. **1**(1), 1–15.

Geronazzo, M., Spagnol, S., Bedin, A., and Avanzini, F. (**2014**). "Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions," in *IEEE Int. Conf. on Acoust., Speech and Sig. Proc. (ICASSP)*, pp. 4496–4500.
Greff, R., and Katz, B. F. G. (**2007**). "Round robin comparison of HRTF simulation results: Preliminary results," in *Audio Eng. Soc. Conv.*, New York, Vol. 123, pp. 2–5.
Guillon, P. (**2009**). "Individualisation des indices spectraux pour la synthèse binaurale: Recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF" ("Individualization of spectral cues for binaural synthesis: Research and exploitation of inter-individual similarities for adaptation of reconstruction of HRTF"), Ph.D. thesis, Université du Maine, Le Mans.
Iwaya, Y. (**2006**). "Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears," Acoust. Sci. Tech. **27**(6), 340–343.
Katz, B. F. G. (**2001a**). "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation," J. Acoust. Soc. Am. **110**(5), 2440–2448.
Katz, B. F. G. (**2001b**). "Boundary element method calculation of individual head-related transfer function. II. Impedance effects and comparisons to real measurements," J. Acoust. Soc. Am. **110**(5), 2449–2455.
Katz, B., and Nicol, R. (**2019**). "Binaural Spatial Reproduction," in *Sensory Evaluation of Sound*, edited by N. Zacharov (CRC Press, Boca Raton), pp. 349–388. ISBN 978-1-4987-5136-0
Katz, B. F. G., and Noisternig, M. (**2014**). "A comparative study of interaural time delay estimation methods," J. Acoust. Soc. Am. **135**(2), 3530–3540.
Katz, B. F. G., and Parseihian, G. (**2012**). "Perceptually based head-related transfer function database optimization," J. Acoust. Soc. Am. **131**(2), EL99–EL105.
Katz, B. F. G., and Picinali, L. (**2011**). "Spatial audio applied to research with the blind," in *Advances in Sound Localization* (IntechOpen, Rijeka), Chap. 13.
Kendall, M. G. (**1938**). "A new measure of rank correlation," Biometrika **30**(1/2), 81–93.
Kendall, M. G. (**1945**). "The treatment of ties in ranking problems," Biometrika **33**(3), 239–251.
Kendall, M. G., and Smith, B. B. (**1939**). "The problem of *m* rankings," Ann. Math. Stat. **10**(3), 275–287.
McMullen, K., Roginska, A., and Wakefield, G. H. (**2012**). "Subjective selection of head-related transfer functions (HRTFs) based on spectral coloration and interaural time differences (ITD) cues," in *Audio Eng. Soc. Conv.*, Vol. 133, pp. 1–9.
Middlebrooks, J. C. (**1999**). "Virtual localization improved by scaling non-individualized external-ear transfer functions in frequency," J. Acoust. Soc. Am. **106**(3), 1493–1510.
Middlebrooks, J. C., Macpherson, E. A., and Onsan, Z. A. (**2000**). "Psychophysical customization of directional transfer functions for virtual sound localization," J. Acoust. Soc. Am. **108**(6), 3088–3091.
Morimoto, M., and Aokata, H. (**1984**). "Localization cues of sound sources in the upper hemisphere," J. Acoust. Soc. Jpn. **5**(3), 165–173.
Parseihian, G., and Katz, B. F. G. (**2012**). "Rapid head-related transfer function adaptation using a virtual auditory environment," J. Acoust. Soc. Am. **131**(4), 2948–2957.
Picinali, L., Afonso, A., Denis, M., and Katz, B. F. (**2014**). "Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge," Int. J. Hum.-Comput. Stud. **72**(4), 393–407.
Poirier-Quinot, D., and Katz, B. F. G. (**2018**). "The Anaglyph binaural audio engine," in *Audio Eng. Soc. Conv.*, Vol. 144, pp. 1–4.
Roginska, A., Wakefield, G. H., and Santoro, T. S. (**2010**). "User selected HRTFs: Reduced complexity and improved perception," Technical Report (Undersea Human Systems Integration Symposium, Providence, RI).
Rumsey, F. (**2012**). *Spatial Audio* (Routledge, Oxford), p. 254.
Schönstein, D., Laurent, F., and Katz, B. F. G. (**2008**). "Comparison of headphones and equalization for virtual auditory source localization," J. Acoust. Soc. Am. **123**, 3724.
Seeber, B. U., and Fastl, H. (**2003**). "Subjective selection of non-individual head-related transfer functions," in *Int. Conf. Auditory Display*, Boston, MA, pp. 259–262.
Simon, L. S. R., Zacharov, N., and Katz, B. F. G. (**2016**). "Perceptual attributes for the comparison of head-related transfer functions," J. Acoust. Soc. Am. **140**(5), 3623–3632.
SOFA. (**2019**). "HRTF database repository," available at https://www.sofaconventions.org/mediawiki/index.php/Files (Last viewed 29 August 2019).
Soper, H. E., Young, A. W., Cave, B. M., Lee, A., and Pearson, K. (**1917**). "On the distribution of the correlation coefficient in small samples. Appendix II to the papers of 'Student' and R. A. Fisher. A cooperative study," Biometrika **11**(4), 328–413.
Stitt, P., Picinali, L., and Katz, B. F. (**2019**). "Auditory accommodation to poorly matched non-individual spectral localization cues through active learning," Sci. Rep. **9**, 1063.
Vigna, S. (**2015**). "A weighted correlation index for rankings with ties," in *Int. Conf. on World Wide Web* (ACM, Florence, Italy), pp. 1166–1176.
Voong, T. M., and Oehler, M. (**2019**). "Tournament formats as method for determining best-fitting HRTF profiles," in *Int. Cong. on Acoust.*, pp. 4841–4847.
Warusfel, O. (**2003**). "IRCAM LISTEN HRTF database," available at http://recherche.ircam.fr/equipes/salles/listen/ (Last viewed 11 July 2019).
Wright, M. (**2005**). "Open sound control: An enabling technology for musical networking," Organised Sound **10**(3), 193–200.
Ziegelwanger, H., Majdak, P., and Kreuzer, W. (**2015**). "Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization," J. Acoust. Soc. Am. **138**(1), 208–222.