OPEN ACCESS Freely available online

C. Kim, V. Lim, and L. Picinali, "Investigation Into Consistency of Subjective and Objective Perceptual Selection of Non-individual Head-Related Transfer Functions" *J. Audio Eng. Soc.*, vol. 68, no. 11, pp. 819–831, (2020 November.).

DOI: https://doi.org/10.17743/jaes.2020.0053

Investigation Into Consistency of Subjective and Objective Perceptual Selection of Non-individual Head-Related Transfer Functions*

CHUNGEUN KIM, *AES Associate*, **VERANIKA LIM, AND LORENZO PICINALI,** *AES Associate*

(ryan.c.kim@gmail.com) (v.lim@imperial.ac.uk) (l.picinali@imperial.ac.uk)

Dyson School of Design Engineering, Imperial College London, United Kingdom

The binaural technique uses a set of direction-dependent filters known as Head-Related Transfer Functions (HRTFs) in order to create 3D soundscapes through a pair of headphones. Although each HRTF is unique to the person it is measured from, due to the cost and complexity of the measurement process pre-measured non-individual HRTFs are generally used. This study investigates whether it is possible for a listener to perceptually select the best-fitting non-individual HRTFs in a consistent manner, using both subjective and objective methods. 16 subjects participated in 3 repeated sessions of binaural listening tests. During each session, participants firstly listened to moving sound sources spatialized using 7 different non-individual HRTFs and ranked them according to perceived plausibility and externalization (subjective selection). They then performed a localization task with sources spatialized using the same HRTFs (objective selection). In the subjective selection, 3 to 9 participants showed test-retest reliability levels that could be regarded as good or excellent depending on the attribute under question, the source type, and the trajectory. The reliability was better for participants with musical training and critical audio listening experience. In the objective selection, it was not possible to find significant differences between the tested HRTFs based on localization-related performances.

0 INTRODUCTION

Throughout the history of research on spatial audio reproduction technology, binaural audio processing has been one of the main techniques under active research, along with loudspeaker-based techniques [1]. Binaural spatial audio rendering makes use of currently available knowledge of the auditory cues at the core of the spatial hearing mechanisms [2]. Among the most widely known are the Interaural Time and Level Differences (ITDs and ILDs, respectively), which provide the sense of lateral source direction. A more comprehensive descriptor that includes the spectral cues as well as ITDs and ILDs has been known as Head-Related Transfer Function (HRTF). HRTFs are the frequency-domain representations of head-related impulse responses (HRIRs), which are measured at the two ears from a given source position, typically in an anechoic environment [3]. Every element of the sound propagation path is therefore comprised in this measurement, including the direction and distance of the source and the morphological characteristics of the head, pinnae, and torso. Rendering of sound sources at arbitrary virtual positions is possible through filtering of the anechoic source signals with the HRTFs measured at the corresponding positions. More convincing simulation of a realistic auditory scene is possible with the additional simulation of reflections or reverberation [4].

The morphological dependence of HRTFs indicates that a set of HRTFs is specific to the person they are measured from. However, it is not practical to collect the HRTF set for every individual listener due to the complexity and cost of the needed equipment and the time for completing the procedure. A common practice, therefore, is to use a generic HRTF set measured with a dummy head produced with morphological dimensions representing a large population [5–7]. There can still be variations between these generic HRTF sets due to the number of dummy head manufacturers and the populations they represent. A number of other HRTF sets, measured on several individuals, are also available as databases from researchers and laboratories worldwide [8–16]. This leads to a pool of non-individual HRTF sets, which can be the starting point of a selection by end users and/or producers of HRTF-based binaural

*Correspondence should be addressed to L. Picinali; e-mail: l.picinali@imperial.ac.uk

audio applications. The next question then remains on how to select an HRTF set from such datasets, which ensures the best performances for a given individual. Various studies have been carried out in the past years to investigate this issue. The focus has been upon the design and validation of methodologies to enable the selection of the best HRTF set or upon a more fundamental issue of the potential success of such methods (for simplicity, the term HRTF will be used in place of HRTF set hereafter to describe a set of HRTF measured from a single person or a dummy head).

Previous HRTF selection attempts can be grouped into two overall categories: physical measurement-based matching and perceptual selection. The former often requires an established HRTF database with the anthropometric measurement of the corresponding individuals, including the head, torso, and various parts of the pinna. For a user whose anthropometric measurement can be taken in the same way, the HRTF that minimizes this measurement difference or the relevant corresponding physical features (such as frequency peaks and notches) can be selected [17–19]. Other similar methods exist where a relationship is established between the HRTF and certain measurements of the pinna (e.g., the outer profile and its distance from the entrance of the ear canal), and then this is used to select the best-fitting HRTF for a given individual from a pre-measured database [20]. This matching or fitting process has increasingly been supported by emerging machine learning techniques [21–24]. The latter (perceptual selection), on the other hand, involves selecting one or multiple HRTFs based on the listener's perception of related spatial or timbral attributes. The scope of this study is specifically upon this perceptual selection procedure, which is not only a separate paradigm under active investigation but is also employed as a validation step in conjunction with many of the physical measurement-based selection techniques.

Previous studies on perceptual selection used strategies that can be further categorized based on the type of test and how the listeners' responses are collected and analyzed. The first kind of these perceptual selection strategies are based on direct subjective rating. The listeners are asked to rate HRTFs based on the perceived quality of some descriptive attributes, from the overall impression [25] to how well the auditory presentation matches specifically described movement or location of the virtual source [26–31]. Various methods have been introduced for the selection, such as ranking [26], rating on scales [27, 28, 30], multiple best choice [29], and pairwise comparisons [25, 31]. Another kind of more indirect perceptual selection strategy is based on objective methods. The listeners perform spatial perception tasks such as sound source localization, from which the selection of the best HRTFs is indirectly made based on the performance, such as the localization errors, front-back confusion rates, and externalization rates [26, 32, 28, 19]. Many of these studies attempted to evaluate their selection strategy and method in terms of the consistency of the selection over repetitions or the performance of the selected HRTFs in a separate validation. However, the general finding from these previous works is that no single best HRTF can be consistently and repeatedly selected. Another interesting finding is that one's own HRTF is "often" found to be the best but not always the single best [25, 32, 27, 28]. Only some marginally successful cases can be found where a set of "better" HRTFs could be chosen [26, 25, 29]. The better-performing HRTFs were relatively easier to outline when the judgments were based on perceived externalization or elevation [29, 19]. Also, grouping of the listeners based on the musical or binaural listening experiences has shown a possibility of better HRTF selectivity by the more experienced ones in that the variances of the judgments were smaller [27]. Other studies employed a prescreening of the listeners, assessing their general ability to localize sound sources within a binaural headphone-based localization task. The best performers are then separated from the worse ones, showing that the selection procedure is consistent for the first and not the latter [20]. Still, these findings were not conclusive, implying the difficulties in understanding the nature of effective HRTF selection.

In a recent study on adaptation to non-individual HRTFs conducted by the authors' research group [33], an attempt was made to pre-select an HRTF suitable for each subject using a localization task within a virtual environment. A preliminary test and analysis showed that there was no consistency in the localization performances (i.e., localization error) using the different HRTFs, making it impossible to assign a single suitable HRTF set per listener. This finding motivated a separate dedicated investigation into the possibility of consistently selecting an HRTF set, preceded by a comprehensive review of the previously used selection methods. This study was therefore initiated aiming at collectively employing the various perceptual HRTF selection methods, found to be potentially effective in the previous works, and assessing the consistency of their results within one set of experiments rather than through separate sets with different listeners. More specifically, a review of the previous studies led to the decision to introduce a subjective selection method of directly grading the HRTFs based on their spatial perceptual attributes and an objective selection method based on the localization accuracy using the stimuli processed with the same HRTFs. The research questions to be answered by this study can be formalized as follows:

- Is there any consistency in listeners' perceptual selection of best HRTFs across repeated sessions, based on subjective or objective methods?
- Is there any effect of previous listening experience on the consistency of the HRTF selection using either methods, as previously found in literature [27, 30]?

The details of the experiment are described in the following sections, followed by the analyses of the results and discussions of the findings.

1 EXPERIMENT

This section describes the design and procedure of the experiment where HRTFs were evaluated and selected repeatedly using the methods introduced in the previous section.

1.1 Participants

A total of 16 participants (11 males and 5 females, aged between 18 and 40) took part in the experiment. All received formal introduction to the procedure using a participant information sheet that, along with the overall experiment protocol, was reviewed and approved by the research ethics committee of the authors' affiliated institution. Written informed consents were obtained from all participants. No auditory or cognitive deficit was reported. In particular, attempts were made to recruit participants with various backgrounds, training, and experience in music and/or 3D audio listening or research, in order to investigate whether the participants with and without experience would show any difference in the results, as the tendency found in [27]. The participants' previous experience in musical training and binaural 3D audio was noted through individual conversations with the experimenter during the introduction. Those with musical training had at least 5 years of experience. Whoever reported to have previous binaural spatial listening experience of any duration were recorded as the ones with binaural audio experience. In [27], a set of criteria was introduced to categorize listeners by mapping their listening experiences to the definitions of generic sensory assessors taken from an ISO standard and adopted to listening tests [34]. According to these criteria, six of the participants who had no musical training or previous experience in binaural spatial audio listening were categorized as initiated assessors (IAs). Four did not have musical training but had had experience in binaural spatial audio listening and were grouped as selected assessors (SAs) using the same criteria. Two were musicians without any binaural audio listening experience and categorized as experts (EXs). Four were musicians or musically trained professionals with extensive experience in spatial audio listening, including binaural synthesis or critical listening skills for audio quality evaluation. These were categorized as expert assessors (EAs). All participants were given a brief tutorial at the beginning of the first session. This consisted in allowing them to carry out both subjective and objective selection tests for a limited amount of time in order to become familiarized with the task and the user interface.

1.2 Experimental Design

The listening test was structured such that the participants were asked to listen to a number of binaural stimuli synthesized using different HRTFs and provide their responses as instructed in three repeated sessions. As described earlier, each session was divided into two sub-sessions, both carried out consecutively on the same day. The subjective selection took an average of 20 minutes across participants, followed by a 10-minute break and the objective selection (40 minutes average). Only one participant took significantly more time (one hour) for each of the sessions. Each session was repeated three times across different days (between two and eleven days apart). The HRTF pool consisted of a total of seven HRTFs that had been selected in [28]. The study performed by Katz and colleagues created a subset of 7 HRTFs (from the 46 included in the IRCAM LISTEN database [10]), perceptually different and optimized in order to satisfy (through a subjective listening test) a large number of participants' judgments. The addition of HRTFs from other databases was initially considered. This would have caused, though, a lengthening of the tests, which would have therefore become more tiring for the participant and potentially less reliable. Furthermore, the addition of HRTFs measured from different labs and with different equipment would have introduced further differences (beyond the actual HRTF individual features) within the set, including asymmetries, spectral magnitude, and ITD variations (see also [35]). This option was therefore discarded.

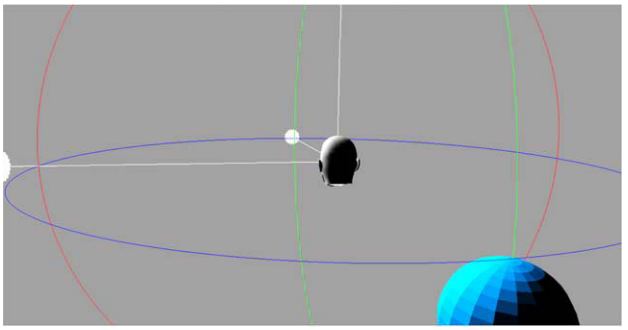
1.2.1 Subjective Selection Method

The participants were firstly asked to perceptually compare and grade the seven HRTFs using discrete scales, focusing specifically on localization-related attributes (i.e., disregarding other attributes such as timbre). Two source signals were used to create the stimuli: an anechoic female speech excerpt from the *Music for Archimedes* collection [36] and bursts of a 200 millisecond-long pink noise with sinusoidal onset and offset ramps of 10 milliseconds. Both source signals were truncated to be 10 seconds long and equalized in the RMS levels. The seven HRTFs were used to simulate continuous circular movement of the sources in two trajectories—horizontal trajectory clockwise starting from the front and vertical trajectory over the sagittal plane starting upward from the front. The speed was adjusted such that one round of circular movement was completed in 10 seconds. No processing was made in terms of distance, which was fixed at the position where the HRTFs were measured [10]. This led to four source signal-trajectory combinations per HRTF.

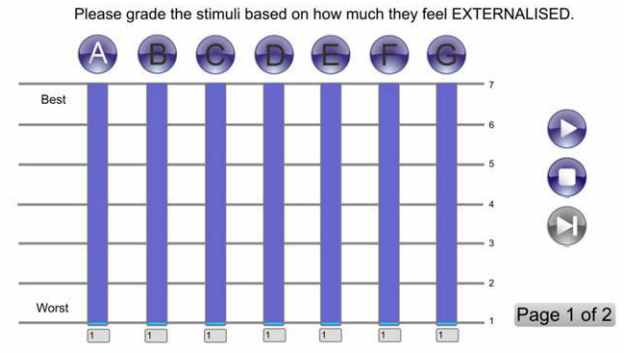
The user interface was developed using Cycling74's MaxMSP, based on an open source patcher originally made for MUSHRA testing.[1] The interface was adapted for the current study, keeping the multi-stimulus playback and switching, and data logging features, and modifying the rest. Each of the source signal-trajectory combinations were presented with the 7 HRTFs on a page with corresponding scales ranging from 1 to 7. The trajectory of the sound source was displayed on a separate screen as the visual reference, as a 3D model of a sphere moving around the head, synchronized in time with the audio. The visual referencing was considered to be an intuitive way of describing the expected source movement at any given moment during the playback. This was done consistently across the participants, since the movement was simulated to be continuous rather than discrete. Any potential bias due to the visual feedback would also be consistent across the HRTFs and repetitions and was therefore not considered to be detrimental to the experiment.

Two localization-based questions were conceived toward the selection based on the reviewed previous studies. The first one was about so-called "plausibility" related to how well the movement of the sound matched the horizontal

[1] https://github.com/IoSR-Surrey/MUSHRA-MaxMSP



(a) Visual representation of moving sound source displayed as a sphere in a third-person viewpoint



(b) Interface on a separate screen, to listen to and rate the seven stimuli using the seven-point discrete scale

Fig. 1. User interface for the subjective HRTF selection test.

or vertical movement of the sphere shown on the separate screen without being skewed, jumping, or giving an impression of front-back or up-down confusions. The second question was about the so-called "externalization," described as how much "outside the head" the participant would feel the sound was moving. Although it could be argued that the sense of plausibility may include externalization, it was decided to collect the grading data for externalization separately, considering that this attribute has been specifically addressed by some of the previous studies reviewed above. Additional questions were initially devised, such as those related with HRTF-describing attributes, as revealed in [37], therefore coloration, immersion, and realism. However, it was decided not to include them all, considering the major increase of the time needed to complete the tests with any additional attribute under question and also considering the comparison of the results with those from the objective selection, purely based on the localization performances. The participants were instructed to try to focus on the two aspects separately, based on the questions asked. A detailed description of the questions was given to the participants both in writing and verbally. The whole set of stimuli was presented twice for these two questions, leading to a total of eight [source signal—trajectory—asked attribute] combinations per HRTF. Consequently, eight pages were presented to each participant for the evaluation of the seven stimuli corresponding to the seven HRTFs. Fig. 1 shows the user interface of one of the evaluation pages with the additional visual reference. The presentation of the pages and

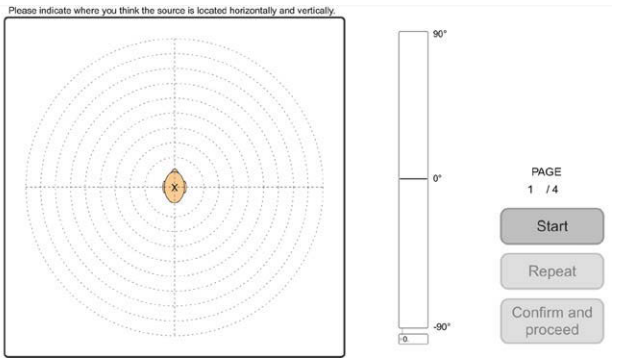


Fig. 2. User interface used for the objective HRTF selection task through localization.

the order of the seven HRTFs per page were randomized in order. It was possible for the participants to control the playback of the stimuli so that they could listen to them or pause as they wanted. It was also possible to switch stimuli seamlessly while they were playing. The participants were asked to use the full range of scales by giving to at least one of the stimuli the highest score (7) and to at least one the lowest (1), while allowed to give the same score to multiple stimuli.

1.2.2 Objective Selection Method

Participants were asked to localize the stimuli presented at fixed positions. The same HRTFs and source signals used in the subjective selection were used with the intention to keep the experimental conditions as similar as possible. However, keeping the signal duration would have made the listening test impractical in terms of the total time. They were therefore truncated to be 4.4 seconds long, corresponding to the length of 1 spoken sentence of the female speech. Six source positions were simulated per source signal and HRTF. These positions were randomized within $\pm 10^\circ$ from 6 fixed azimuth-elevation pairs, $(0^\circ, 10^\circ)$, $(180^\circ, 10^\circ)$, $(60^\circ, 40^\circ)$, $(-60^\circ, 40^\circ)$, $(120^\circ, -20^\circ)$, and $(-120^\circ, -20^\circ)$, to prevent potential confinement of presented source directions when completely randomized. Similarly to the subjective selection, no processing was made in terms of distance, which was fixed at the position where the HRTFs were measured [10]. This led to a total of 84 stimuli to be evaluated (2 source signals, 7 HRTFs, and 6 virtual source positions).

The user interface was developed using the same software used for the subjective selection. Fig. 2 shows the user interface designed for this experiment. On each page, one of the stimuli was presented, and the participant was asked to indicate the perceived source position on a circular horizontal plane and separate vertical scale. The azimuth and distance could be marked on the horizontal plane and the elevation on the vertical scale. The horizontal plane was set to have a maximum radial distance of approximately 2 m from the center of the head, which is in line with the distance at which the HRTFs were measured. Instead of direct marking of the distance numerically, a diagram of

the head was displayed in the center, which implied how far the source could be distributed. Because the absence of any additional reverberation could result in the source being localized inside the head, the participants were reminded that they could mark the perceived position of the source inside the head drawn in the middle of the diagram as well as in other positions within the plane. They were also asked to maintain the same criteria across the various trials. Considering that no simulation was performed for rendering sources at different distances, this metric was used to enable collection of data that could be related to the perception of externalization and compared to the rating of the HRTFs from the subjective selection experiment. The main point of interest here was the difference between the HRTF sets and the consistency in the evaluations rather than the validation of the distance indication method. The participants were allowed to repeat each stimulus before confirming their answer and proceeding to the next.

1.3 Listening Environment and Equipment

The test was conducted in a dedicated quiet room, acoustically treated with absorbent foam. The participants were given a laptop with the user interface, connected to an additional larger monitor. The audio was played using an external USB audio interface (MOTU UltraLite Mk3) through Sennheiser HD 650 open-back headphones. The spatialized audio was generated using the 3D Tune-In Toolkit Test Application [38], which was controlled by the Max interface using the Open Sound Control (OSC) protocol. No head-tracking was used in order to keep the experimental conditions in line with the ones used in previous studies. Participants were encouraged not to move their heads during the experiment.

The frequency response of the audio rendering chain was measured using a calibrated microphone and frequency analyzer (NTI XL2 with M2230 microphone), and the prominent uneven responses with over ± 3 dB differences from the flat lower frequency region were compensated by using a linear phase parametric equalizer (MOTU CueMix 7-band EQ, with phase-lock) on the output, mimicking the inverse response. The output level was initially adjustable by each participant around a previously fixed level in case it was not comfortable. After this initial calibration, it then remained the same throughout the experiment per participant. The applied gain ranged from -3 dB to +9 dB from the originally fixed level with the majority of the participant remaining at the original level.

Previous studies have shown that the transfer function between headphones and ear drums (HpTF) can play a role in terms of externalization and overall naturalness of the binaural rendering [39, 40]. Nevertheless, strong evidence does not exist to support that HpTFs can improve localization accuracy [41, 42]. Furthermore, HpTFs are not direction dependent, as they are applied in the same way to all source positions, and do not therefore have an influence on HRTF-specific effects, which are the objects of this study. For these reasons, and in line with relevant research in the area (e.g., [30]), no HpTF was measured and used in this study.

2 RESULTS AND ANALYSES

The results from the experiments over the three repeated sessions are outlined and analyzed in this section. The two experimental sessions corresponding to the subjective and objective HRTF selection methods are analyzed separately.

2.1 Subjective Selection

The results for the subjective selection experiment are presented in the following subsection.

2.1.1 Test-Retest Reliability as a Measure of Consistency

The Intra Class Correlation (ICC) was used as the indicator of the test-retest reliability [43–45], which in turn will work as a measure of selection consistency. Individual ICC estimates (per participant) and their 95% confidence intervals were calculated using the SPSS statistical analysis package version 24, based on a mean-rating ($k = 3$), absolute-agreement, 2-way mixed-effects model [45]. This corresponds to the nature of the experimental design involving a single rater and the three repeated sessions of rating seven HRTFs per one trial.

Based on a 95% confidence interval, Koo and Li [45] give the following recommendation for interpreting ICC:

- below 0.50: poor
- between 0.50 and 0.75: moderate
- between 0.75 and 0.90: good
- above 0.90: excellent

These values were then separately derived for the specific cases of each independent variable—attribute under question (plausibility or externalization), source signal (speech or noise), and trajectory (horizontal or vertical). Table 1 shows the ICC values for all participants, across all trials per session and per each of the test conditions.

2.1.2 Reliability Levels for All and Specific Test Conditions

Across all the test conditions, three of the participants in the expert assessor (EA) group and one in the initiated assessor (IA) group show good overall reliability level.

The individual reliability levels vary when the specific test conditions are separately examined. When the ICC values are compared only for the question on the externalization, three participants in the EA group, two in the SA group, three in the IA group, and one from the EX group show a good reliability level. Of these, the one EX has the highest ICC (0.894). On the question of plausibility, two participants in the EA group and one in the SA group show a good or higher reliability level. A reliability level above 0.9 (“excellent”) is seen with the EA participant. Looking at the type of sound sources, for the noise, three participants in the EA group show good reliability levels. With

Table 1. Intra Class Correlation (ICC) estimates for the three repeated sessions in the subjective HRTF selection. The values are calculated per participant, firstly across all case combinations and then for each case of the independent variables (attribute under question, source type, and trajectory). The last column on the right contains the data for the single combination of independent variables that gave overall highest ICC estimates (Externalization-Speech-Vertical). The underlined participants belong to the expert assessors group (EA) and the ICC values over 0.75 ("good" reliability or higher) are marked in **bold**.

	Overall	Question		Source type		Trajectory		Ext-Sp-Ver
		Extern	Plaus	Noise	Speech	Horz	Vert	
Participant 1 (IA)	0.317	0.468	<0.001	<0.001	0.527	0.163	0.286	0.630
Participant 2 (IA)	0.422	0.416	0.429	0.194	0.574	<0.001	0.697	**0.923**
Participant 3 (SA)	0.537	**0.760**	0.184	0.235	0.685	0.270	0.710	**0.955**
Participant 4 (SA)	0.737	0.690	**0.779**	0.656	**0.805**	0.693	**0.765**	0.682
Participant 5 (EA)	**0.871**	**0.888**	**0.858**	**0.861**	**0.885**	**0.813**	**0.923**	**0.918**
Participant 6 (SA)	<0.001	<0.001	0.283	0.089	<0.001	0.086	<0.001	<0.001
Participant 7 (EA)	0.342	<0.001	0.582	0.397	0.221	0.449	0.230	<0.001
Participant 8 (EX)	0.399	0.376	0.427	0.349	0.451	0.347	0.452	0.285
Participant 9 (IA)	**0.782**	**0.827**	0.737	0.737	**0.823**	**0.773**	**0.798**	**0.809**
Participant 10 (SA)	0.642	**0.769**	0.466	0.690	0.609	0.685	0.541	0.650
Participant 11 (IA)	0.303	<0.001	0.459	0.315	0.261	0.355	0.218	0.057
Participant 12 (EA)	**0.874**	**0.797**	**0.938**	**0.882**	**0.870**	**0.893**	**0.857**	**0.752**
Participant 13 (IA)	0.523	**0.767**	0.004	0.487	0.568	0.467	0.581	**0.939**
Participant 14 (IA)	0.713	**0.877**	0.401	0.533	**0.830**	**0.773**	0.652	**0.939**
Participant 15 (EA)	**0.751**	**0.771**	0.736	**0.847**	0.637	**0.754**	**0.754**	**0.773**
Participant 16 (EX)	0.690	**0.894**	0.183	0.605	**0.768**	0.643	0.739	**0.952**

the speech as the source, two participants in the EA group, one from the SA group, two from the IA group, and one from the EX group show good reliability levels. Lastly, for the horizontal source trajectory, three participants in the EA group and two in the IA group show good reliability levels. For the vertical source trajectory, three participants in the EA group, one in the SA group, and one in the IA group show good or higher reliability levels. A reliability level above 0.9 ("excellent") is seen with an EA participant. In general, the highest ICC values are seen with the EAs, except for the question on externalization (one EX had the highest ICC). Only two participants (5 and 12, both EAs) show a good or excellent reliability level for every individual test condition. Interestingly, one participant from the EA group (Participant 7) does not show a reliability level above 0.75 in any test condition. On the other hand, one participant (Participant 9) from the IA group shows overall high reliability levels for all the test conditions, compared to the other IAs; the smallest ICC estimate for this participant (0.737) is near the "good" range.

2.1.3 Differences in Reliability Between Expertise Levels

The estimated ICC values for all the participants and each case of the questions asked, source trajectories, and signal types were examined to find out whether there was any difference in the reliability levels between the participants with a higher level of expertise and the rest. The participants were divided into two groups—those who were categorized as EA (Participants 5, 7, 12 and 15), with the higher expected level of expertise, and those in the other categories (IA, SA, and EX). An independent-samples Mann-Whitney U test was conducted due to the non-normality of the data sets. For the plausibility question, the participants labeled as experts showed higher test-retest reliability ($M = 0.78$, $SD = 0.16$) than the participants labeled as non-experts ($M = 0.37$, $SD = 0.23$) for all trials ($U = 44$, $p = 0.013$). Also for all trials with the noise as the source, the participants labeled as experts showed higher reliability ($M = 0.75$, $SD = 0.23$) than the non-experts ($M = 0.40$, $SD = 0.26$; $U = 42$, $p = 0.030$). No significant difference was found between the two groups for the overall reliability level or for the other trial conditions.

2.1.4 Difference Between Condition Pairs

In order to compare the ICC values for all the participants between each pair of conditions given in the experiment—question type (externalization or plausibility), source type (noise or speech), and trajectory (horizontal or vertical)—Wilcoxon Signed Ranks Tests were conducted. There was no significant difference in the ICC values between the externalization ($M = 0.48$, $SD = 0.64$) and plausibility ($M = 0.48$, $SD = 0.28$) questions ($Z = -0.595$, $p = 0.552$), the noise ($M = 0.48$, $SD = 0.29$) and speech ($M = 0.55$, $SD = 0.40$) source types ($Z = -1.16$, $p = 0.245$), or the horizontal ($M = 0.50$, $SD = 0.30$) and vertical ($M = 0.53$, $SD = 0.40$) trajectories ($Z = -0.170$, $p = 0.865$).

2.1.5 Independent Variables Combination

An attempt was made to find the combination across the various independent variables, which resulted in the highest ICC estimates across all participants. The combination of Externalization (question), Speech (source type), and Vertical (trajectory) led to the overall highest ICC values ($M = 0.64$, $SD = 0.34$), with 9 participants achieving ICC above the 0.75 threshold (good or excellent reliability level). Wilcoxon Signed Ranks Tests were conducted in order to verify whether the differences between this combination of variables and the other possible ones were significantly different, but the results outlined that this was not the case.

2.1.6 Summary

The findings from the subjective selection experiment can be summarized as follows. Firstly, 4 participants out of the 16 showed overall test-retest reliability levels (across all the test questions, source types, and trajectories) higher than 0.75, therefore at the "good" level or above. Among the four were three of the participants from the expert assessor (EA) group. For individual test conditions, there were case-dependent variations in the number and composition of participants who showed good and higher levels of reliability. The number of participants with good or excellent reliability levels was the smallest at three for the test trials with plausibility as the question and noise as the source, and the largest at nine for the trials with externalization as the questioned attribute. In addition to all these individual differences, the comparison between the EA group and the rest of the participants showed statistically significant reliability differences in two specific test cases—the plausibility question and noise source type. A combination of the various independent variables was found with the highest ICC estimates (Externalization-Speech-Vertical), but the differences between this and the other possible combinations were not statistically significant.

2.2 Objective Selection

The results for the objective selection experiment are presented in the following subsections. These were analyzed with the intention of assessing consistency in the evaluation of the HRTFs in terms of sound localization performances rather than by subjectively rating them as in the previous section. In order to do so, we first had to establish whether the best HRTF per participant could be robustly determined and/or whether the HRTFs could be ranked as in the subjective selection, based on the participant's localization results. Localization errors and front-back confusion rates were compared with the rating of the the plausibility question in the subjective selection, and perceived distances were compared with the externalization question. Raw data were used without pre-processing except for an overall initial analysis to identify major issues (e.g., errors in the data collection system, etc.). Participants showing very large localization errors were not discarded as outliers, as this would have acted against one of the purposes of the study (i.e., assessing the consistency of non-individual HRTF rating with expert and non-expert listeners).

2.2.1 Localization Angular Errors

The localization errors were calculated in three different forms: the overall localization angle errors and the lateral and intraconic angle errors [46]. The lateral coordinate represents the angle of the source from the median plane (from -90° to $+90^\circ$) and the intraconic coordinate represents rotation around the interaural axis from the horizontal plane (from 0° to 360°). This separation of the overall error was made with the intention to observe whether there was any trend in the consistency specifically in terms of the lateralization, reflecting the ability to utilize ITD and ILD, or in terms of the elevation judgment, reflecting the use of the spectral characteristics of the sound. When calculating the intraconic angle errors, all target and response locations were projected onto the frontal hemisphere in order to exclude the effect of front-back confusion leading to excessively large error values. Front-back confusion rates were analyzed separately.

A non-parametric Kruskal-Wallis test was performed to verify whether there was any significant difference between the HRTFs per participant for each source type (speech or noise), each session (first, second, or third), and the three types of localization errors. Statistically significant differences were found with six of the participants, only for specific source types and specific error types. Table 2 shows for which participants and in which test conditions significant differences between the HRTFs were found. Fig. 3 shows the box plots for these participants and for the corresponding localization measures and test conditions. There was no participant who showed significant difference between the HRTFs in more than one session for the same source type. This implies that even if some significant distinction between the HRTF might have been possible, this was not repeated across the different sessions, making further consistency investigations (e.g., toward the ranking of HRTFs) irrelevant.

2.2.2 Front-Back Confusion Rates

The front-back confusion rates were calculated as the ratio between the number of trials where the confusion is observed and the total trials per session and per HRTF for each participant. This led to a single confusion rate value per HRTF per session derived from all of the corresponding trials. The distribution of the front-back confusion rates across the HRTFs showed no noticeable tendency or pattern over the three sessions. For a few participants, a significant difference (Kruskal-Wallis) could be found between the HRTFs for a specific session, and it was therefore possible to select one HRTF as the best one (i.e., the one with the lowest front-back confusion rate). Fig. 4 shows the example case of Participant 7. Table 3 lists the participants for which a significant difference could be found between HRTFs for a given session, reporting the HRTF identified as the best one. It can be noted that the best HRTFs per participant are not the same for both source types and when the source types are disregarded, the best HRTFs are even less identifiable.

2.2.3 Perceived Distance

Similarly to the localization error measures, a non-parametric Kruskal-Wallis test was also performed to see whether there were significant differences between the HRTFs for each participant at each session in terms of perceived distance. As listed in Table 2, only Participant 3 showed statistically significant difference between the HRTFs, only for the speech source and the first session. Therefore, also with this metric, it was not possible to find significant differences between the HRTFs and across the sessions.

Table 2. Combinations of test conditions (source type, session number, and localization measure) for the seven participants for whom statistically significant differences between the HRTFs were found. The participants in the expert assessors (EA) group are underlined.

	Source type	Session no.	Localization measure	Kruskal-Wallis test Sig.
Participant 3 (SA)	Speech	First	Perceived distance	0.034
Participant 4 (SA)	Noise	Second	Intraconic angle error	0.017
Participant 8 (EX)	Speech	Third	Total angle error	0.025
	Noise	Third	Intraconic angle error	0.002
Participant 9 (IA)	Noise	Third	Intraconic angle error	0.009
Participant 11 (IA)	Speech	Second	Intraconic angle error	0.019
Participant 15 (EA)	Noise	First	Intraconic angle error	0.014

Table 3. List of participants for whom the best HRTF was identified in all the three repeated sessions in terms of the front-back confusion rate, for each of the source types (first two columns) and across both source types (third column). N/C denotes "Not Conclusive." The participants in the EA group are underlined.

	Best HRTF index for:		
	Speech	Noise	All
Participant 1 (IA)	3	N/C	N/C
Participant 4 (SA)	N/C	3	3
Participant 5 (EA)	6	1	N/C
Participant 6 (SA)	N/C	6	N/C
Participant 7 (EA)	7	3	N/C
Participant 8 (EX)	7	N/C	N/C
Participant 9 (IA)	N/C	2	N/C
Participant 11 (IA)	N/C	N/C	6
Participant 13 (IA)	N/C	7	2

2.2.4 Additional Test-Retest Reliability Check

As described previously, the results from the Kruskal-Wallis tests implied that even before considering consistency over session repetitions, the HRTFs could not be distinguished from each other based on the localization errors. However, for comparison of the equivalent parameters, the ICC was calculated as well for the intraconic angle errors over the three sessions as an example, in order to allow for a comparison with the subjective selection. In order to conduct the ICC analysis, 252 trials were aggregated to generate means for each unique condition, i.e., session (3 levels), source type (2 levels), and HRTF (7 levels). Based on the results from the previous analyses, we expected low ICCs across sessions for the majority of the participants.

Overall, only Participant 4 showed an ICC estimate of 0.654, at the "moderate" level, between the three sessions with a 95% confidence interval from 0.117 to 0.881 ($F(13,26) = 2.752$, $p = 0.014$). No other significant estimate of ICC was found, implying that no consistency could be concluded in the overall HRTF selection based on the elevation errors across the sessions. Looking separately at the two sound source types, considering the speech source type, no test-retest reliability was found between the three sessions for any participant. However, considering the noise, ICC estimates at the "good" level were found between the three sessions for Participants 4 and 6. For Participant 4, the average measure ICC was 0.810 with a 95% confidence interval from 0.323 to 0.964 ($F(6,12) = 5.167$, $p = 008$). For Participant 6, the average measure ICC was 0.804 with a 95% confidence interval from 0.349 to 0.962 ($F(6,12) = 5.778$, $p = 005$).

The high variability of the confidence intervals in these two cases, and the fact that for all other participants the ICC estimates were far below the 0.75 threshold used for the subjective selection experiment, indicate that no conclusion can be drawn from the ICC results, as expected. Considering these results, and the ones obtained previously, no further analysis was carried out for the other localization error metrics.

2.2.5 Summary

The findings from the objective selection experiment can be summarized as follows: when the localization results

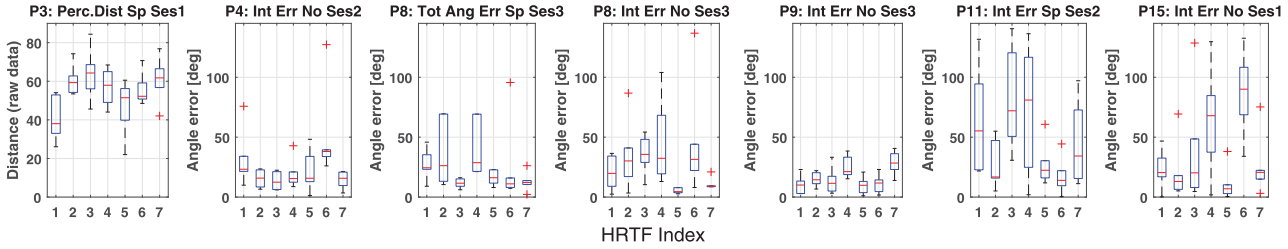


Fig. 3. Box plots (median and the 25th and 75th percentiles, with outliers marked as '+') of localization measures for the participants listed in Table 2, for whom statistically significant differences were found between the HRTFs in the corresponding test conditions. P denotes Participant and Ses denotes Session number. Note that the reported distance is expressed as raw data, where 100 is the maximum allowed by the interface. Despite the statistically significant differences between the HRTFs, these were only specific to single sessions and no further tendency was found that would lead to any consistency across the three sessions.

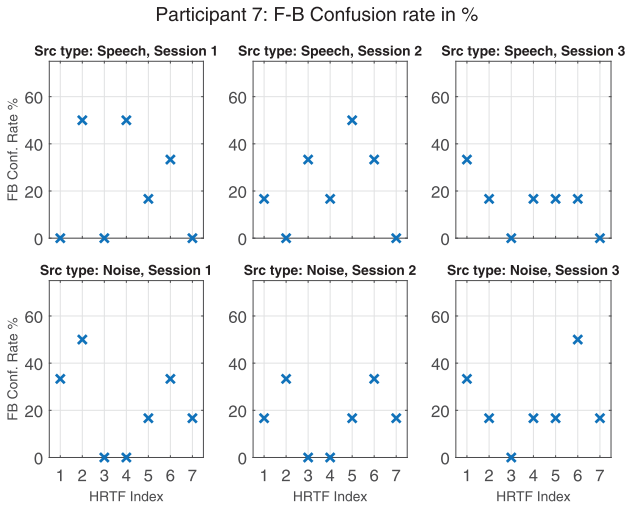


Fig. 4. Front-back confusion rates for Participant 7 across sessions. For the speech and noise as the source, HRTF numbered 7 and 3 respectively show the lowest confusion rates across all the three sessions.

were collected and analyzed aiming at ranking the various HRTFs, distinction between the HRTFs was possible with six of the participants only for certain source types and localization measures and only in single specific sessions. The analysis of the front-back confusion rates did not show any significant difference between the HRTFs, which was consistent across the repetitions, nor helped in identifying a single best HRTF per participant over all the test conditions. Consequently, further IIC-based consistency analysis from the ranking of HRTFs by means of localization accuracy, toward equivalent comparison to the subjective selection, was not possible.

3 DISCUSSION

3.1 Subjective Selection

The first general finding from the results is that "good-or-above" consistency in the HRTF selection across the different sessions is only observed with a few of the participants. Only two participants showed the ICC values of good or excellent range overall and in all the individual test conditions. This is in line with previous studies introduced earlier, which at best showed that a certain number of "better" HRTFs could be selected repeatedly by a small number of listeners among the tested ones (e.g., [30]). Variations in the source type, trajectory, and the attribute under question did not seem to help improve the consistency. Another noticeable finding, although not entirely conclusive, is the relatively better consistency found in the expert assessor (EA) group, again in line with previous literature. The distinction between this EA group and the rest was statistically significant specifically with plausibility as the question and the noise as the source type. This listener categorization was based on the works of [27, 34] and the tendency is in agreement with what was seen in [27]. Nevertheless, the presence of non-EAs also with high ICC estimates (in some test conditions) suggests that a more detailed survey of the participants' previous listening experience would have been beneficial. It is important to underline that this study did not only aim to assess the consistency in the identification of one best (or worse) HRTF across different testing sessions but also (and mainly) to look at the consistency of the overall ranking of the selection. The assumption that a listener should be able to rank non-individual HRTFs in a similar way across different sessions, assuming that the assessment method and conditions are the same, is a valid one and has also been explored in previous research [30].

3.2 Objective Selection

On the other hand, the results from the objective selection experiment did not reveal any meaningful tendency related to the selection consistency. This was mainly due to the fact that the distinction between the HRTFs within a single session was not possible based on a localization accuracy metric. This is in line with some of the previous similar studies introduced earlier [26, 32]. It was however not possible to observe any correspondence of the results to those of the subjective selection, such as in the investigation of [28]. This is probably due to the fact that in that study the HRTFs used for the objective evaluation had already been pre-selected using a subjective method, whereas in our study no pre-selection was made, with the intention of having a bias-free comparison of the two selection methods. Another element that could be considered in order to explain these results is that in the current experiment the user interface for localization was only two-dimensional while other similar studies employed 3D pointing methods, such as that of [28]. Furthermore, the scale of the head diagram on the horizontal plane compared to the whole usable range implies that whenever the virtual sources were perceived inside the head, the resolution for the localization indication might not have been sufficient. However, the findings are not different from those of our preliminary investigation, in which a mobile phone-based head-tracked virtual environment was used for source orientation indication [33]. Considering that two-dimensional interfaces have already been employed as pointing methods in localization tasks (e.g., [47]), it seems unlikely that the difference in the user interface would have resulted in any unexpected deviation of the findings. These findings are only in partial agreement with previous works on the subject and they add on to the doubt that a method based on rapid localization-based non-individual HRTF selection tests might not be robust and repeatable enough in order to obtain consistent results.

3.3 HRTFs Perception Training

Considering the results of the subjective selection experiment, and in particular those related with expert assessors, a hypothesis can be formulated about the potential effect of training on the consistency of HRTFs perceptual ratings. Studies have shown how listeners can adapt to non-individual HRTFs through repeated training sessions; the measured outcome used to quantify the adaptation has, though, very often been a measure of localization accuracy [48–51]. Investigating the effect of training on subjective

metrics (e.g., the consistency of the ranking of a set of non-individual HRTFs across different sessions) could allow us to better understand the extent of the auditory accommodation mechanisms and support the design of novel HRTF training procedures and techniques.

3.4 Implications in HRTF Individualization

The findings of this study, along with the many others on HRTF selection techniques, suggest that perceptual selection of non-individual HRTFs based on subjective/qualitative evaluations is possible to a certain extent only when using expert assessors. On the other hand, consistent selection based on objective evaluations still represents an open challenge, and research is not currently leading to positive results.

Looking back at the method and procedures used in the two experiments, the following considerations can be made:

- In both tests, listeners were not allowed to interact with the sound sources (e.g., no head-tracking was allowed). This would not have worked for the sound localization task but it might have been beneficial for the subjective selection task.
- The set of the seven HRTFs employed in both experiments was selected in a study from Katz and colleagues [28] using a subjective method, very similar to the one employed here. It might be the case that such an HRTFs sub-set cannot be considered representative of the tested population and therefore using additional HRTFs (possibly from other open datasets) could lead to more positive results in both objective and subjective selection tasks.
- Related to the point above, in the current study we did not have any participant performing the two experiments with their own measured HRTF. This option would have represented a "control" condition to verify the robustness of the method and test procedure and to be used as a baseline for the analysis of the results of both objective and subjective selection.
- The fact that the test was repeated across several days might have added unexpected differences in the presentation of the rendered audio due to the variations in the headphone placement across sessions.
- As mentioned earlier, exploring alternative interfaces for the user to report the location of the source in the objective selection task could potentially lead to different results (see also [52]).

It is therefore possible to envisage that further research investigating some of these matters more in depth might lead to the development of a method that will result in more conclusive findings in terms of perceptual-based HRTF selection.

4 CONCLUSION

This study was devised with a view to better comprehend the selectivity of non-individual HRTFs with attempts to incorporate some of the previously used perception-based selection strategies. Two widely used approaches were tested—subjective selection, where the listeners directly graded HRTFs based on perceptual attributes, and objective selection, where various localization accuracy measures were derived and used for HRTF rating. In the subjective selection, many of the participants were found to have generally poor or moderate test-retest reliability levels, whereas some, mostly from the group who had previous binaural audio experiences, did show good or excellent levels of reliability. This implied potential influences of prior musical training combined with experience in binaural audio on the consistency of direct HRTF selection, which is partially in line with previous literature on the subject. In the objective selection, examination of all the extracted localization-related measures based on the participants' trials revealed that it was not possible to statistically distinguish the HRTFs over the repeated test sessions. Although a few "better" HRTFs could be identified for some participants, no clear tendency was found to be able to support selectivity of a single non-individual HRTF per participant for all the tested conditions. This indicated that the localization-based indirect selection is even more challenging if compared with subjective ranking. Overall, the difficulty of perceptual selection suggests not only that other selection methods should be investigated further but also that auditory training and other cognitive processes need to be incorporated in such research. It is important to underline the limitations of this study, mainly related with methodological and experimental choices (as can be seen in the previous section), which should be taken into account when attempting to generalize these findings.

5 ACKNOWLEDGMENT

This study was supported by the 3D Tune-In project, European Union's Horizon 2020 research and innovation program under grant agreement No. 644051. The authors are grateful to Peter Stitt, David Thery, and Brian FG Katz for their feedback in drafting the manuscript.

6 REFERENCES

[1] F. Rumsey, *Spatial Audio* (Focal Press, Oxford, 2001).

[2] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. (Brill, Leiden, The Netherlands, 2013).

[3] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display*, 2nd ed. (J. Ross Publishing, 2013).

[4] D. Begault, "3-D Sound for Virtual Reality and Multimedia," *Tech. rep.*, NASA Ames Research Center (2000).

[5] M. D. Burkhard and R. M. Sachs, "Anthropometric Manikin for Acoustic Research," *J. Acoust. Soc. Am.* (1975), https://doi.org/10.1121/1.380648.

[6] K. Genuit, H. W. Gierlich, and W. Bray, "Development and Use of Binaural Recording Technique," presented at the *89th Convention of the Audio Engineering Society* (1990 Sept.), convention paper 2950.

[7] F. Christensen, C. B. Jensen, and H. Møller, "The Design of VALDEMAR-An Artificial Head for Binaural Recording Purposes," presented at the *109th Convention of the Audio Engineering Society* (2000 Sept.), convention paper 5253.

[8] J. Blauert, M. Brueggen, A. W. Bronkhorst, R. Drullman, G. Reynaud, L. Pellieux, W. Krebber, and R. Sottek, "The AUDIS Catalog of Human HRTFs," *J. Acoust. Soc. Am.*, vol. 103, no. 5, p. 3082 (1998), https://doi.org/10.1121/1.422910.

[9] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF Database," *Proc. 2001 IEEE Workshop Appl. Signal Process. Audio Acoust. (Cat. No.01TH8575)*, pp. 99–102 (2012), https://doi.org/10.1109/ASPAA.2001.969552.

[10] O. Warufsel, "IRCAM LISTEN HRTF Database," http://recherche.ircam.fr/equipes/salles/listen/ (2002).

[11] N. Gupta, A. Barreto, M. Joshi, and J. C. Agudelo, "HRTF Database at FIU DSP Lab," *2010 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 169–172 (2010), https://doi.org/10.1109/ICASSP.2010.5496084.

[12] T. Nishino, N. Inoue, K. Takeda, and F. Itakura, "Estimation of HRTFs on the Horizontal Plane Using Physical Features," *Appl. Acoust.*, vol. 68, no. 8, pp. 897–908 (2007), https://doi.org/10.1016/j.apacoust.2006.12.010.

[13] B. Xie, X. Zhong, D. Rao, and Z. Liang, "Head-Related Transfer Function Database and its Analyses," *Sci. China Series G: Phys. Mech. Astron.*, vol. 50, no. 3, pp. 267–280 (2007), https://doi.org/10.1007/s11433-007-0018-x.

[14] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," presented at the *German Annual Conference on Acoustics (DAGA)* (2013).

[15] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database," *Appl. Sci. (Switzerland)* (2018), https://10.3390/app8112029.

[16] ARI HRTF Database - Acoustics Research Institute (ARI), Austrian Academy of Sciences, http://www.kfs.oeaw.ac.at/hrtf.

[17] D. Y. N. Zotkin, J. Hwang, R. Duraiswaini, L. S. Davis, "HRTF personalization using anthropometric measurements," presented at the *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, pp. 157–160 (2003).

[18] K. Iida, Y. Ishii, and S. Nishioka, "Personalization of Head-Related Transfer Functions in the Median Plane Based on the Anthropometry of the Listener's Pinnae," *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 317–333 (2014), https://doi.org/10.1121/1.4880856.

[19] M. Geronazzo, S. Spagnol, A. Bedin, and F. Avanzini, "Enhancing Vertical Localization With Image-Guided Selection of Non-Individual Head-Related Transfer Functions," *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 4463–4467 (2014), https://doi.org/10.1109/ICASSP.2014.6854446.

[20] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini, "Applying a Single-Notch Metric to Image-Guided Head-Related Transfer Function Selection for Improved Vertical Localization," *J. Audio Eng. Soc.*, vol. 67, no. 6, pp. 414–428 (2019 Jun.).

[21] S.-N. Yao, T. Collins, and C. Liang, "Head-Related Transfer Function Selection Using Neural Networks," *Archives Acoust.*, vol. 42, no. 3, pp. 365–373 (2017), https://doi.org/10.1515/aoa-2017-0038.

[22] K. Yamamoto and T. Igarashi, "Fully Perceptual-Based 3D Spatial Sound Individualization With an Adaptive Variational Autoencoder," presented at the *ACM Transactions on Graphics* (2017), https://doi.org/10.1145/3130800.3130838.

[23] T. Y. Chen, T. H. Kuo, and T. S. Chi, "Autoencoding HRTFS for DNN Based HRTF Personalization Using Anthropometric Features," presented at the *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (2019), https://doi.org/10.1109/ICASSP.2019.8683814.

[24] G. W. Lee and H. K. Kim, "Personalized HRTF Modeling Based on Deep Neural Network Using Anthropometric Measurements and Images of the Ear," *Appl. Sci. (Switzerland)* (2018), https://doi.org/10.3390/app8112180.

[25] S. Yairi, Y. Iwaya, and Y. Suzuki, "Individualization Feature of Head-Related Transfer Functions Based on Subjective Evaluation," presented at the *14th International Conference on Auditory Display* (2008).

[26] B. U. Seeber and H. Fastl, "Subjective Selection of Non-Individual Head-Related Transfer Functions," *2003 Int. Conf. Audit. Display*, pp. 259–262 (2003).

[27] D. Schönstein and B. F. G. Katz, "Variability in Perceptual Evaluation of HRTFs," *J. Audio Eng. Soc*, vol. 60, no. 10, pp. 783–793 (2012 Nov.).

[28] B. F. G. Katz and G. Parseihian, "Perceptually Based Head-Related Transfer Function Database Optimization," *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. EL99–EL105 (2012), https://doi.org/10.1121/1.3672641.

[29] Y. Wan, A. Zare, and K. McMullen, "Evaluating the Consistency of Subjectively Selected Head-Related Transfer Functions (HRTFs) Over Time," presented at the *AES 55th International Conference: Spatial Audio*, pp. 1–8 (2014 Aug.), conference paper 5-5.

[30] A. Andreopoulou and B. Katz, "Investigation on Subjective HRTF Rating Repeatability," presented at the *140th Convention of the Audio Engineering Society* (2016 May), convention paper 9597.

[31] R. Shukla, R. Stewart, A. Roginska, and M. Sandler, "User Selection of Optimal HRTF Sets via Holistic Comparative Evaluation," presented at the *2018 AES International Conference on Audio for Virtual and Augmented Reality (AVAR)* (2018 Aug.), conference paper P4-2.

[32] A. Härmä, R. van Dinther, T. Svedström, M. Park, and J. Koppens, "Personalization of Headphone Spatialization Based on the Relative Localization Error in an Auditory Gaming Interface," presented at the *132nd Convention of the Audio Engineering Society* (2012 Apr.), convention paper 8644.

[33] C. Kim, M. Steadman, J. -H. Lestang, D. F. M. Goodman, and L. Picinali, "A VR-Based Mobile Platform for Training to Non-Individualized Binaural 3D Audio,"

presented at the *144th Convention of the Audio Engineering Society* (2018 May), convention paper 10010.

[34] N. Zacharov and G. Lorho, "What Are the Requirements of a Listening Panel for Evaluating Spatial Audio Quality?" presented at the *Spatial Audio and Sensory Evaluation Techniques Workshop* (2006).

[35] A. Andreopoulou, D. R. Begault, and B. F. Katz, "Inter-Laboratory Round Robin HRTF Measurement Comparison," *IEEE J. Selected Topics Signal Process.*, vol. 9, no. 5, pp. 895–906 (2015).

[36] Bang and Olufsen , "Music for Archimedes" (1992).

[37] L. S. R. Simon, N. Zacharov, and B. F. G. Katz, "Perceptual Attributes for the Comparison of Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 140, no. 5, pp. 3623–3632 (2016), https://doi.org/10.1121/1.4966115.

[38] M. Cuevas-Rodriguez, L. Picinali, D. Gonzalez-Toledo, C. Garres, E. de la rubia Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona, "3D Tune-In Toolkit: An Open-Source Library for Real-Time Binaural Spatialisation," *PLoS ONE*, vol. 14, no. 3 (2019).

[39] N. I. Durlach, A. Rigopulos, X. Pang, W. Woods, A. Kulkarni, H. Colburn, and E. Wenzel, "On the Externalization of Auditory Images," *Presence: Teleop. Virtual Env.*, vol. 1, no. 2, pp. 251–257 (1992).

[40] B. Masiero and J. Fels, "Perceptually Robust Headphone Equalization for Binaural Reproduction," presented at the *130th Convention of the Audio Engineering Society* (2011 May), convention paper 8388.

[41] I. Engel, D. L. Alon, P. W. Robinson, and R. Mehra, "The Effect of Generic Headphone Compensation on Binaural Renderings," presented at the *2019 AES International Conference on Immersive and Interactive Audio* (2019 Mar.), conference paper 73.

[42] D. Schonstein, L. Ferré, and B. F. Katz, "Comparison of Headphones and Equalization for Virtual Auditory Source Localization," *J. Acoust. Soc. Am.*, vol. 123, no. 5, p. 3724 (2008).

[43] P. E. Shrout and J. L. Fleiss, "Intraclass Correlations: Uses in Assessing Rater Reliability," *Psych. Bull.*, vol. 86, no. 2, pp. 420–428 (1979).

[44] J. P. Weir, "Quantifying Test-Retest Reliability Using the Intraclass Correlation Coefficient and the SEM," *J. Strength Condition. Res.*, vol. 19, no. 1, p. 231 (2005).

[45] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *J. Chiro. Med.*, vol. 15, no. 2, pp. 155–163 (2016), https://doi.org/10.1016/j.jcm.2016.02.012.

[46] M. Morimoto and H. Aokata, "Localization Cues of Sound Sources in the Upper Hemisphere," *J. Acoust. Soc. Japan (E)*, vol. 5, no. 3, pp. 165–173 (1984).

[47] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916 (2001).

[48] P. Stitt, L. Picinali, and B. F. Katz, "Auditory Accommodation to Poorly Matched Non-Individual Spectral Localization Cues Through Active Learning," *Sci. Rep.*, vol. 9, no. 1 (2019).

[49] M. A. Steadman, C. Kim, J. -H. Lestang, D. F. Goodman, and L. Picinali, "Short-Term Effects of Sound Localization Training in Virtual Reality," *Sci. Rep.*, vol. 9, no. 1, pp. 1–17 (2019).

[50] E. Fuchs and G. Flügge, "Adult Neuroplasticity: More Than 40 Years of Research," *Neural Plast.*, vol. 2014, pp. 1–10 (2014), https://doi.org/10.1155/2014/541870.

[51] S. Carlile and T. Blackman, "Relearning Auditory Spectral Cues for Locations Inside and Outside the Visual Field," *JARO - J. Assoc. Res. Otol.*, vol. 15, no. 2, pp. 249–263 (2014).

[52] H. Bahu, T. Carpentier, M. Noisternig, and O. Warusfel, "Comparison of Different Egocentric Pointing Methods for 3D Sound Localization Experiments," *Acta Acust. united Acust.*, vol. 102, no. 1, pp. 107–118 (2001).

THE AUTHORS


Chungeun Kim


Veranika Lim


Lorenzo Picinali

Chungeun Kim received his BSc in Electrical Engineering at Seoul National University, MSc in Sound and Vibration Studies at Institute of Sound and Vibration, University of Southampton, and PhD in Psychoacoustical Engineering at Institute of Sound Recording, University of Surrey. Since 2009, he has worked as a postdoctoral researcher at University of Surrey, Eindhoven University of Technology, and Imperial College London in a number of research projects. His research areas include auditory perception and quality evaluation, computational modeling of auditory processes, and binaural audio. He currently works as a senior engineer at Qualcomm Technologies International Ltd.

•

Veranika Lim has a degree in Cognitive Psychology and Cognitive Neuroscience from Leiden University in the Netherlands, with a research internship at the Flight Deck Display Research Lab (FDDRL), NASA, USA. She holds a PhD degree from Eindhoven University of Technology (Netherlands) for her thesis on 'Design Opportunities in Reducing Domestic Food Waste: A Collective Approach.' Afterward, she joined the Audio Experience Design group within the Dyson School of Design Engineering at Imperial College London as a Research Associate collaborating on the H2020 cultural heritage project named PLUGGY. Veranika is currently working at *yulife*, a well-being and life insurance company, doing product research and design.

•

Lorenzo Picinali is a Senior Lecturer in Audio Experience Design at Imperial College London. In the past years he worked in Italy, France, and the UK on projects related with 3D binaural sound rendering, interactive applications for visually and hearing impaired individuals, audiology and hearing aids technology, audio and haptic interaction, and more in general acoustical virtual and augmented reality. He is currently co-leading the design and implementation of the 3D Tune-In Toolkit, an open-source C++ library for audio spatialization and simulation of hearing loss and hearing aids.