Audio Engineering Society

Conference Paper

Presented at the International Conference on
Headphone Technology
2019 August 27–29, San Francisco, CA, USA

Perceptual Evaluation of Personalized BRIRs and Headphone Compensation

Graham Davis[1], Andre Schevciw[1], Isaac Munoz[1], and Nils Peters[1]

[1] *Qualcomm Technologies, Inc. San Diego, CA, USA*

Correspondence should be addressed to Graham Davis (grahamd@qti.qualcomm.com)

ABSTRACT

Previous literature demonstrates two key components required for the accurate reproduction of a 3-dimensional sound field over headphones: head-related transfer functions (HRTFs) and headphone compensation filters (HCFs). This study seeks to supplement an already extensive body of work with a new methodology for the capture and evaluation of individualized binaural room impulse responses (BRIRs) and HCFs. The results of a double-blind listening test corroborate earlier findings regarding the interaction of BRIR and HCF source on binaural audio quality. Finally, the perceptual advantages of a fully individualized binaural rendering procedure are reported.

1. INTRODUCTION

Increasing commercialization of mixed reality technology, hearable devices, and immersive audio renderers necessitates the transparent and spatially accurate reproduction of recorded and synthesized sound fields over headphones. There exists a vast corpus of prior research exploring the techniques involved in generating effective binaural content for such purposes [1]. In the work that follows, we consider the binaural room impulse response (BRIR) and headphone compensation filter (HCF) as two key components of the binauralization process.

The human auditory system relies on a variety of perceptual cues (e.g. coloration, interaural time differences, and interaural level differences) to determine the distance and direction of a sound source [2]. Individual anatomical structures encode this information in the pressure wave of an external sound source as it is reflected and diffracted off and around the pinna, head, and torso [1]. The HRTF and its inverse Fourier Transform (the head-related impulse response [HRIR]) represent these directional cues as a set of linear filters. Every individual has a unique set of HRIRs that can be measured or computed. HRIRs derived in a reverberant environment are termed BRIRs. Previous papers note the importance of HRIR and BRIR individualization with regards to the perceptual authenticity and angular accuracy of binaural audio [1][3][4][5].

Beyond personalized HRIR and BRIR measurements, one must also consider the binaural reproduction system. Schärer & Lindau explored the perceptual quality of several headphone equalization techniques [6]. Using an ABC with hidden reference listening test design, subjects were asked to rate the perceived difference between binaural test signals and a reference loudspeaker. Six test conditions incorporated varying HCFs, and the seventh forwent headphone compensation altogether. The results of this study demonstrated the perceptual disadvantage of uncompensated binaural content. Moreover, listeners reported spectral qualities such as high

frequency ringing, timbral difference, and poor bass as key differentiators between the reference loudspeaker and headphone reproduction [6].

Brinkmann employed a similar listening test to examine the impact of individual versus non-individual HCFs [7]. Unexpectedly, non-individual compensation generated from the FABIAN head and torso simulator (HATS) was rated higher than individually compensated content. Furthermore, 72% of participants noted spectral coloration as a key attribute of dissimilarity between the loudspeaker reference and binaural reproduction [7].

The two previously cited studies rely on non-individual HRTFs. This investigation builds upon such results with the inclusion of a fully individualized binauralization process. In what follows, we develop a procedure for the capture and assessment of individualized BRIRs and HCFs. First, a system for the repeatable measurement of BRIRs and headphone impulse responses (HPIRs) is introduced, followed by an inverse filter design method for the generation of HCFs. Next, a listening test methodology is proposed for the qualitative evaluation of binaural reproduction. Finally, results from the experiment are presented along with a discussion of their impact.

2. BRIR & HPIR MEASUREMENTS

All system identification is accomplished through Farina's logarithmic swept sine technique [8] with a sample rate of 48kHz. Automated measurement procedures are implemented in MATLAB and incorporate third party software for audio I/O [9] and impulse response deconvolution [10].

BRIRs and HPIRs are commonly measured using the blocked ear canal method [7]. Because the ear canal acts as a transmission line independent of source direction, measurements made at the entrance of the ear canal contain the encoded spatial information that our auditory system relies on to localize sound [2]. Figure 1 depicts a custom in-ear coupler used for the repeatable placement of DPA4060 microphones at the entrance of a blocked ear canal.

Figure 1. In-ear microphone holder for BRIR and HPIR measurements.

Variability of in-ear microphone placement is measured with a GRAS KEMAR HATS in an anechoic chamber. The frontal HRTF is measured ten times with the DPA4060 microphones removed and replaced between measurements. Variability is quantified as the standard deviation of each FFT frequency bin between repeated measurements. As is demonstrated in Figure 2, variability never exceeds 2dB in the frequency range 100Hz-20kHz.

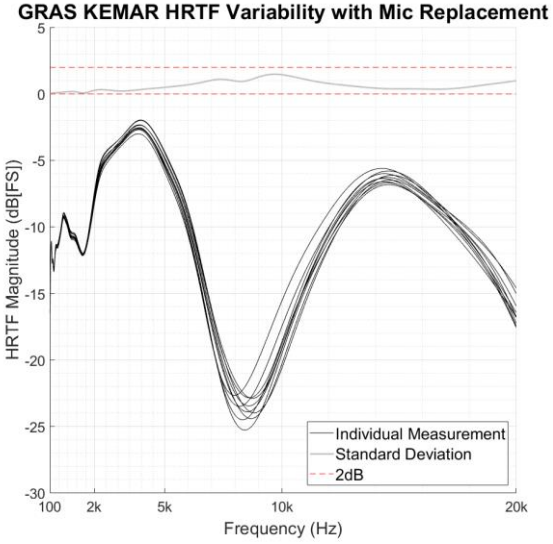

Figure 2. Ten repeated 6th octave band smoothed GRAS KEMAR frontal HRTFs and their standard deviation per FFT frequency bin.

2.1 BRIR Measurement

In this study, BRIRs are measured in a critical listening room equipped with full-range Genelec 8250A loudspeakers calibrated using Genelec's GLM 3.0 AutoCal software. The room dimensions conform to Recommendation ITU-R BS.1116-3 [11] as outlined in Table 1. The BRIR from a speaker location is calculated as the mean of four repeated swept sine measurements. The resulting BRIR is truncated and windowed to 8192 samples (~170ms) to contain the entire audible reverberation of the test environment. A Tukey window with a taper ratio of 0.7 is used for windowing to ensure a continuous taper to zero on both ends of the BRIR.

Length	6.8m
Width	4.2m
Height	5.0m
Aspect ratio $1.1 \cdot w/h \leq l/h$ fulfilled	Yes
Aspect ratio $l/h < 3$ fulfilled	Yes
Aspect ratio $w/h < 3$ fulfilled	Yes
RT60	0.16s

Table 1. Room dimensions and properties

2.2 HPIR Measurement

HPIR measurements are made in the critical listening room described in Section 2.1. Open-backed Sennheiser HD800 headphones are used for HPIR measurement and binaural playback throughout the remainder of this study. The HPIR for each ear is calculated as the mean of five repeated swept sine measurements, with the headphones removed and replaced between each measurement. The mean HPIR is truncated and windowed to 2048 samples (~43ms).

2.2.1 Intra-Individual HPIR Variability

Previous research demonstrates intra-individual variance of HPIRs, particularly around high frequency notches [6][7][12][13]. Figure 3 presents the magnitude response variability of 5 repeated HPIR measurements for a single subject. The standard deviation between repeats is negligible below 2kHz. However, 6-10dB peaks in standard deviation are seen around HPTF notch frequencies at 8.5, 12.5, and 21kHz. The depth and center frequency of HPTF notches vary as the headphones are removed and replaced between repeats. Consequently, direct inversion of a single HPTF measurement may result in a compensation filter with misplaced and/or exaggerated peaks. Such peaks are reported to result in audible ringing artifacts [7]. Therefore, averaging of HPIRs and regularized inversion (discussed in Section 3) are employed to mitigate the consequences of intra-individual HPIR variability.

3. COMPENSATION FILTER DESIGN

Headphone compensation filters are generated from the HPIRs measured in Section 2.2. The filter design procedure is accomplished in MATLAB, leveraging the AKregulatedInversion method in the AKTools MATLAB toolbox [10]. Minimum-phase, frequency domain, least mean squares (LMS) inversion is achieved using Equation 1, as presented by Brinkmann [7] and developed by Norcross et al. [14].

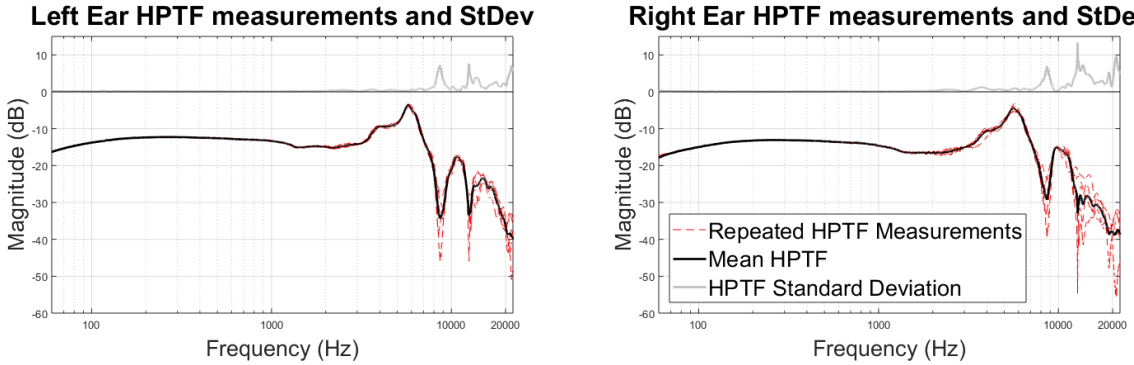

Figure 3. Five repeated Sennheiser HD800 HPTF measurements, mean HPTF, and HPTF standard deviation for one subject. Peaks in standard deviation are seen around notch frequencies (8.5, 12.5 and 21kHz).

$$H_c'(\omega) = \frac{HPTF^*(\omega)A(\omega)}{|HPTF(\omega)|^2} \quad (1)$$

Where, $HPTF^*(\omega)$ is the complex conjugate of the measured HPTF. A target bandpass filter, $D(\omega)$, and regularization parameters are contained within the magnitude response of $A(\omega)$.

$$|A(\omega)| = \frac{|D(\omega)|}{1+\beta\frac{|B(\omega)|^2}{|HPTF(\omega)|^2}} \quad (2)$$

In the current study, $|D(\omega)|$ is composed of a 2nd order Butterworth high pass with 100Hz cutoff in cascade with a 1st order Butterworth lowpass with 20kHz cutoff. A 6th octave band smoothed direct inversion of $HPTF(\omega)$ is used as the regularization spectrum, $|B(\omega)|$. The regularization weight is set as $\beta = 0.2$. Furthermore, $|B(\omega)|$ is constrained to have a flat response below 1kHz and above 20kHz. As explained in [7] and discussed in Section 2.2.1, regularization minimizes the compensation of notches in the measured HPTF, mitigating unwanted sharp peaks in the resulting compensation filter.

A minimum phase target response is obtained via the Hilbert transform (see Equation 3).

$$\angle A(\omega) = -\text{imag}(\text{Hilbert}(\ln(|A(\omega)|))) \quad (3)$$

The resulting compensation spectrum, $H_c'(\omega)$, is smoothed by 1/6th octave bands before taking an IFFT to retrieve the time-domain FIR compensation filter. Figure 4 shows a sample left ear HPTF, the regularization curve used for inversion, and the final compensation filter spectrum (HCF).

4. ASSESSMENT

The qualitative assessment methodology defined in this section consists of two measurement phases and one double-blind listening test. All phases are completed in one session in the critical listening room described in Section 2.1. The entire assessment session lasts approximately 45 minutes per subject.

The first measurement phase consists of individual HPIR measurement (as described in Section 2.2) and HCF design (as described in Section 3). The second phase consists of individual BRIR measurement (as described in Section 2.1) from six discrete speaker directions of a 28.2 loudspeaker layout presented in Table 2. The two measurement phases are completed with a GRAS KEMAR HATS for generic test conditions. The third phase is a listening test, described in the following section. The subject is asked to sit comfortably, facing straight ahead for the completion of all three phases.

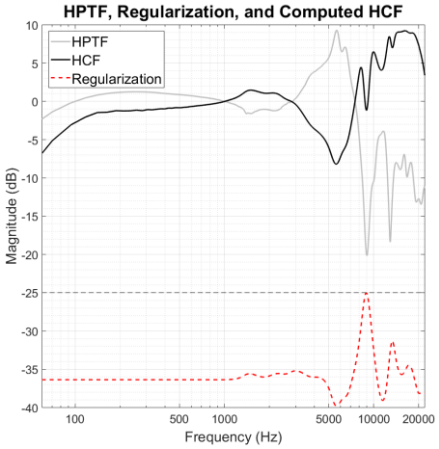

Figure 4. Measured HPTF, HCF, and regularization curve for sample subject's left ear. Regularization curve magnitude offset by -25dB for visualization.

Code	Spk #	Azi	Ele
F (front)	1	+000	+000
FL (front-left)	2	+030	+000
L (left)	6	+090	+000
RR (rear-right)	11	-135	+000
FLU (front-left-up)	14	+045	+035
FRD (front-right-down)	28	-045	-015

Table 2. BRIR directions used in listening test.

4.1 Listening Test Methodology

4.1.1 Interface

The user interface depicted in Figure 5 was developed in Python. It consists of three rating scales (one for each of the qualitative attributes described in Section 4.1.2), a reference playback control, and a test source playback control. No limit is set on the number of times a subject may play either the reference or test source. Subjects are required to set a score for every qualitative attribute before moving to the next trial. The interface is projected on a screen in front of the subject and is controlled with a wireless mouse.

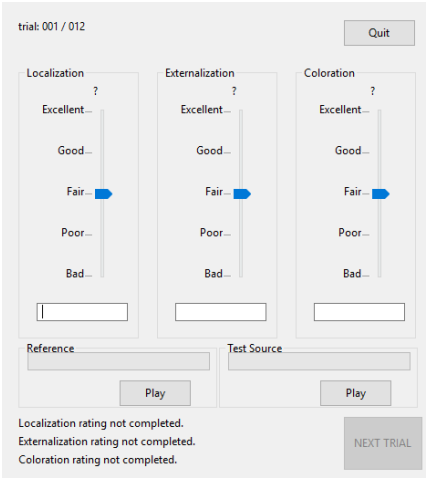

Figure 5. Listening test graphical user interface.

4.1.2 Evaluation Parameters

Subjects are asked to evaluate every test source along three attributes; localization, externalization, and coloration. An integer scale from 1 to 5 (labelled Bad, Poor, Fair, Good, and Excellent) is used to compare the test source presented over headphones to the loudspeaker reference. Evaluation instructions and rating scale verbiage are projected alongside the test interface and reproduced in Table 3 for convenience.

Comparing the headphone presentation with the loudspeaker presentation, how do you judge the headphone presentation with regards to the following qualitative attributes?		
Coloration		
5	Excellent	(source coloration matches reference)
4	Good	(source coloration nearly matches reference)
3	Fair	(source coloration differs somewhat from reference)
2	Poor	(source coloration differs substantially from reference)
1	Bad	(source coloration differs extremely from reference)
Localization		
5	Excellent	(source azimuth and elevation match reference)
4	Good	(source azimuth and elevation nearly match reference)
3	Fair	(source azimuth and elevation differ somewhat from reference)
2	Poor	(source azimuth and elevation differ substantially from reference)
1	Bad	(source azimuth and elevation differ extremely from reference)
Externalization		
5	Excellent	(source externalization and distance match reference)
4	Good	(source externalization nearly matches reference)
3	Fair	(source externalization differs somewhat from reference)
2	Poor	(source externalization differs substantially from reference)
1	Bad	(source externalization differs extremely from reference)

Table 3. Assessment question and rating scale presented to subjects during the listening test.

4.1.3 Stimuli

An order 15 maximum length sequence (MLS) with duration 0.7 seconds is chosen as the base stimulus for binaural evaluation. This decision is informed by Brinkmann's finding that subjects can differentiate binauralized noise from a reference loudspeaker more easily than a binauralized natural sound source [7].

Each trial's test source is played over headphones and is generated through time domain convolution of the MLS base signal with one BRIR, with or without headphone compensation. The full matrix of test conditions is presented in Table 4. The reference signal is the MLS played from the loudspeaker corresponding to that trial's BRIR direction. Subjects are asked to remove the headphones while listening to the reference signal.

	No HCF	**Individual HCF**	**KEMAR HCF**
Individual BRIR	C01	C02	C03
KEMAR BRIR	C04	C05	C06

Table 4. Test condition matrix.

Beyond the 36 test stimuli (six test conditions from each of the six BRIR speaker directions), subjects are presented with six anchor conditions found in Table 5. Anchor conditions are generated using C02 from the frontal BRIR direction. There are two anchor conditions targeting each of the three qualities described in Section 4.1.2. A01 and A02 target localization error through rotation of the sound field. A03 and A04 demonstrate reduced externalization through truncation of the BRIR reverberant response. Finally, A05 and A06 present degradations in test source coloration through spectral filtering.

Condition	**Rotation**	**BRIR Length**	**Lowpass**
A01 (loc1)	90º	8192	None
A02 (loc3)	30º	8192	None
A03 (ext1)	0º	256	None
A04 (ext3)	0º	2048	None
A05 (col1)	0º	8192	3.5kHz
A06 (col3)	0º	8192	7kHz

Table 5. Definition of anchor conditions.

4.1.4 Training Session

Each subject completes a training session before the listening test to familiarize themselves with the test interface and task. The training session consists of seven trials, each with a different condition (C02 and the six anchor conditions). Training conditions are chosen to demonstrate the range of possible test conditions. Training trials are presented in a random order from the frontal BRIR speaker location. Data from the training session is not included in analysis. The training session lasts approximately 5-10 minutes and the test administrator is present in the room to answer any procedural questions.

4.1.5 Test Design

The listening test consists of 42 randomized trials, including the six anchor conditions from Table 5 and 36 test conditions (each of the six conditions illustrated in Table 4 presented from the six speaker locations in Table 2).

Each trial lasts roughly 30 seconds, for a total of approximately 20 minutes. Subjects were given a five-minute break after the first 21 trials. At the end of the listening test, subjects participate in a short debrief session with the test administrator.

5. RESULTS

25 subjects (5 female and 20 male) participated in the binaural evaluation described in Section 4.1. Participants were all Qualcomm employees ranging from 25 to 60 years of age.

Mean opinion scores (MOS), standard errors (SE), and 95% confidence intervals (CI95) for the six conditions are presented in Figure 6 and Table 6.
For individual BRIR conditions (C01, C02, and C03), we see the following trend in MOS for all qualities:

Individual HCF > Generic HCF > No HCF

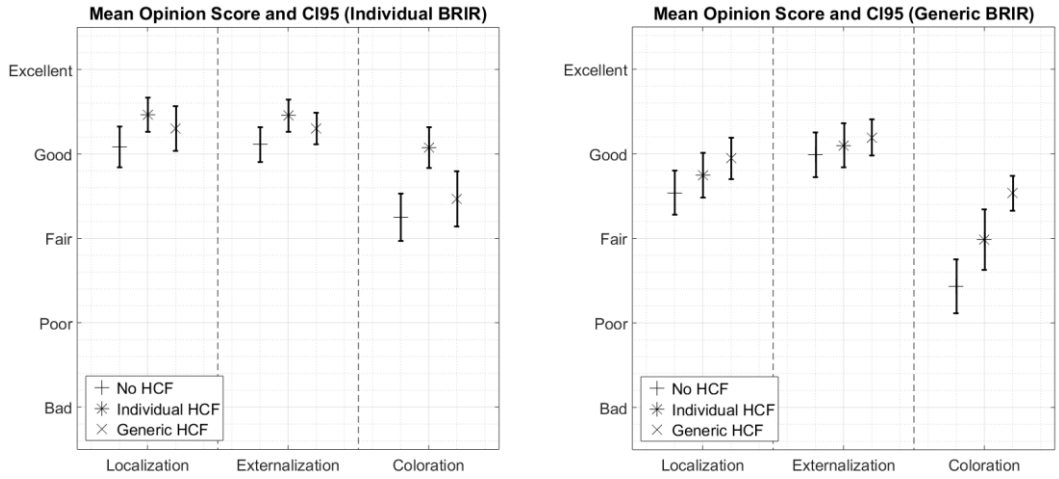

Figure 6. MOS results for all six conditions including student t-distribution CI95 with N=25 (t[24] = 2.064).

Test Conditions			Mean			Standard Deviation			CI95% (+/-)		
Condition	**BRIR**	**HCF**	**LOC**	**EXT**	**COL**	**LOC**	**EXT**	**COL**	**LOC**	**EXT**	**COL**
C01	Individual	None	4.08	4.11	3.25	0.59	0.50	0.68	0.24	0.21	0.28
C02	Individual	Individual	4.47	4.45	4.07	0.49	0.46	0.58	0.20	0.19	0.24
C03	Individual	Generic	4.30	4.30	3.47	0.64	0.45	0.79	0.26	0.18	0.33
C04	Generic	None	3.54	3.99	2.43	0.64	0.64	0.78	0.26	0.27	0.32
C05	Generic	Individual	3.75	4.10	2.99	0.63	0.64	0.87	0.26	0.26	0.36
C06	Generic	Generic	3.95	4.19	3.53	0.60	0.52	0.50	0.25	0.22	0.21

Table 6. MOS results for all six conditions including student t-distribution CI95 with N=25 (t[24] = 2.064).

Similarly, a trend in MOS is seen for generic BRIR conditions (C04, C05, and C06):

Generic HCF > Individual HCF > No HCF

C02 has a numerically higher MOS than all other conditions for all qualities. C04 has the lowest MOS for every quality. In what follows, we explore the statistically significant interactions of our dependent variables; BRIR source, HCF source, and condition.

5.1 Results by BRIR Source

A one-way ANOVA (α=0.05) of results grouped by BRIR source was carried out in MATLAB. BRIR source significantly influences localization MOS ($F[1, 48] = 11.89$, $p = 0.001$) and coloration MOS ($F[1, 48] = 13.12$, $p = 0.001$). Individual BRIRs have statistically higher MOS for both localization ($M = 4.28$, $SD = 0.53$) and coloration ($M = 3.60$, $SD = 0.56$) than generic BRIRs. Although individual BRIRs ($M = 4.29$, $SD = 0.42$) have a higher MOS for externalization, the improvement over a generic BRIR ($M = 4.09$, $SD = 0.57$) is not statistically significant ($F[1, 48] = 1.92$, $p = 0.17$).

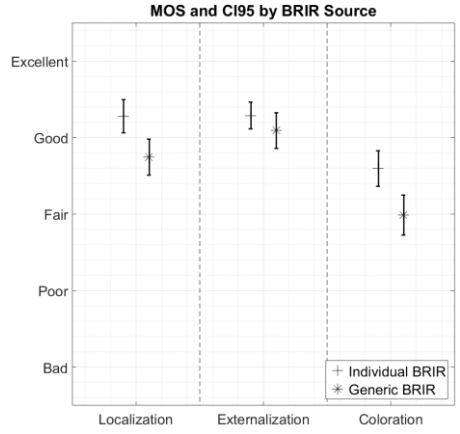

Figure 7. MOS and CI95 by BRIR source.

	BRIR	Mean	StDev	SE	CI95%
Localization	**Individual**	4.28	0.53	0.11	0.22
	Generic	3.74	0.58	0.12	0.24
Externalization	**Individual**	4.29	0.42	0.08	0.18
	Generic	4.09	0.57	0.11	0.23
Coloration	**Individual**	3.60	0.56	0.11	0.23
	Generic	2.98	0.63	0.13	0.26

Table 7. Descriptive statistics by BRIR source.

5.2 Results by HCF Source

A one-way ANOVA (α=0.05) of results grouped by HCF source was carried out in MATLAB. HCF source demonstrates a significant impact on coloration MOS ($F[2, 72] = 9.72$, $p < 0.001$). The impact of HCF source on localization ($F[2, 72] = 2.74$, $p = 0.07$) and externalization MOS ($F[2, 72] = 1.51$, $p = 0.23$) is not statistically significant.

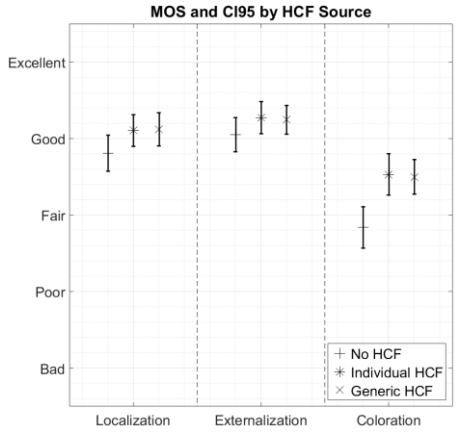

Figure 8. MOS and CI95 by HCF source.

	HCF	Mean	StDev	SE	CI95%
Localization	**None**	3.81	0.57	0.11	0.23
	Individual	4.11	0.50	0.10	0.21
	Generic	4.12	0.53	0.11	0.22
Externalization	**None**	4.05	0.53	0.11	0.22
	Individual	4.28	0.51	0.10	0.21
	Generic	4.25	0.46	0.09	0.19
Coloration	**None**	2.84	0.66	0.13	0.27
	Individual	3.53	0.66	0.13	0.27
	Generic	3.50	0.55	0.11	0.23

Table 8. Descriptive statistics by HCF source.

				Confidence Interval	
HCF One	**HCF Two**	**p**	**Estimate**	**Lower**	**Upper**
None	**Individual**	**0.001****	**-0.690**	**-1.113**	**-0.267**
None	**Generic**	**0.001****	**-0.660**	**-1.083**	**-0.237**
Individual	Generic	0.984	0.030	-0.393	0.453
Bold row = significant at * = 95% CI, ** = 99% CI					

Table 9. HCF coloration post hoc test results.

Post hoc comparisons of coloration MOS by HCF source was accomplished in MATLAB with a Tukey's Honest Significant Difference (HSD) test of multiple means [15]. Two significant interactions can be seen in the test results presented in Table 9. Both individual and generic HCF demonstrate an improvement over no compensation.

5.3 Results by Condition

A one-way ANOVA ($\alpha = 0.05$) of results grouped by condition (as presented in Figure 6 and Table 6) was carried out in MATLAB. A statistically significant interaction between conditions exists for localization ($F[5, 144] = 8.22$, $p < 0.001$), externalization ($F[5, 144] = 2.33$, $p = 0.045$), and coloration ($F[5, 144] = 15.10$, $p < 0.001$).

Post hoc comparisons of localization, externalization, and coloration MOS were accomplished in MATLAB with a Tukey's HSD test of multiple means. Statistically significant interactions for each quality are reported. If not listed, it can be assumed that the trend between two conditions is not statistically significant.

5.3.1 Localization

Table 10 presents post hoc test results for localization MOS by condition. C04 MOS (M = 3.54, SD = 0.60) is lower than all individual BRIR conditions (C01, C02, C03). Furthermore, C02 MOS (M = 4.47, SD = 0.49) is higher than all generic BRIR conditions. Finally, C03 MOS (M = 4.30, SD = 0.64) is higher than C05 (M = 3.75, SD = 0.63).

Condition One	Condition Two	p	Estimate	Confidence Interval	
				Lower	Upper
C01	C02	0.203	-0.387	-0.870	0.097
C01	C03	0.787	-0.220	-0.704	0.264
C01	**C04**	**0.018****	**0.540**	**0.056**	**1.024**
C01	C05	0.363	0.333	-0.150	0.817
C01	C06	0.970	0.133	-0.350	0.617
C02	C03	0.924	0.167	-0.317	0.650
C02	**C04**	**0.000****	**0.927**	**0.443**	**1.410**
C02	**C05**	**0.000****	**0.720**	**0.236**	**1.204**
C02	**C06**	**0.027***	**0.520**	**0.036**	**1.004**
C03	**C04**	**0.000****	**0.760**	**0.276**	**1.244**
C03	**C05**	**0.014***	**0.553**	**0.070**	**1.037**
C03	C06	0.297	0.353	-0.130	0.837
C04	C05	0.828	-0.207	-0.690	0.277
C04	C06	0.157	-0.407	-0.890	0.077
C05	C06	0.847	-0.200	-0.684	0.284
Bold row = significant at * = 95% CI, ** = 99% CI					

Table 10. LOC post hoc test results by condition.

5.3.2 Externalization

Table 11 presents post hoc test results for externalization MOS by condition. C02 (M = 4.45, SD = 0.46) has a significantly higher externalization rating than C04 (M = 3.99, SD = 0.64).

Condition One	Condition Two	p	Estimate	Confidence Interval	
				Lower	Upper
C01	C02	0.228	-0.340	-0.776	0.096
C01	C03	0.828	-0.187	-0.623	0.250
C01	C04	0.963	0.127	-0.310	0.563
C01	C05	1.000	0.013	-0.423	0.450
C01	C06	0.995	-0.080	-0.516	0.356
C02	C03	0.918	0.153	-0.283	0.590
C02	**C04**	**0.028***	**0.467**	**0.030**	**0.903**
C02	C05	0.191	0.353	-0.083	0.790
C02	C06	0.533	0.260	-0.176	0.696
C03	C04	0.316	0.313	-0.123	0.750
C03	C05	0.782	0.200	-0.236	0.636
C03	C06	0.982	0.107	-0.330	0.543
C04	C05	0.977	-0.113	-0.550	0.323
C04	C06	0.757	-0.207	-0.643	0.230
C05	C06	0.990	-0.093	-0.530	0.343
Bold row = significant at * = 95% CI, ** = 99% CI					

Table 11. EXT post hoc test results by condition.

5.3.3 Coloration

Table 12 presents the test results for coloration by condition. As with localization, C04 (M = 2.43, SD = 0.78) is rated lower than all individual BRIR conditions. Furthermore, C04 is rated lower than C06 (M = 3.53, SD = 0.50). C02 (M = 4.07, SD = 0.58) receives a higher coloration MOS than all conditions except for C06.

Condition One	Condition Two	p	Estimate	Confidence Interval	
				Lower	Upper
C01	**C02**	**0.001****	**-0.827**	**-1.401**	**-0.253**
C01	C03	0.885	-0.220	-0.794	0.354
C01	**C04**	**0.001****	**0.813**	**0.239**	**1.387**
C01	C05	0.790	0.260	-0.314	0.834
C01	C06	0.713	-0.287	-0.861	0.287
C02	**C03**	**0.031***	**0.607**	**0.033**	**1.181**
C02	**C04**	**0.000****	**1.640**	**1.066**	**2.214**
C02	**C05**	**0.000****	**1.087**	**0.513**	**1.661**
C02	C06	0.079	0.540	-0.034	1.114
C03	**C04**	**0.000****	**1.033**	**0.459**	**1.607**
C03	C05	0.162	0.480	-0.094	1.054
C03	C06	0.999	-0.067	-0.641	0.507
C04	C05	0.066	-0.553	-1.127	0.021
C04	**C06**	**0.000****	**-1.100**	**-1.674**	**-0.526**
C05	C06	0.073	-0.547	-1.121	0.027
Bold row = significant at * = 95% CI, ** = 99% CI					

Table 12. COL post hoc test results by condition.

6. DISCUSSION

The results presented in this study corroborate the conclusions of prior research and support the a priori hypothesis that a fully individualized binauralization procedure (C02) is the optimal method for accurate spatial and spectral reproduction of immersive audio over headphones.

In Section 5.1, we demonstrated the perceptual benefits of individualized BRIRs over generic BRIRs. As Møller et al. [5] report a decrease in localization errors with individual binaural recordings, our findings reveal a significant improvement in subjective localization ratings with the use of personalized BRIRs. Furthermore, the perceptual relevance of individual pinna and torso related spectral features discussed in Xie [1] are supported by our improved coloration ratings for individualized BRIRs. Although not statistically significant, our results also demonstrate a trend towards improved externalization through BRIR individualization.

Schärer and Lindau [6] establish the importance of headphone compensation on transparent binaural reproduction. In Section 5.2, we verified the significant benefits of individual and generic headphone compensation on binaural coloration. Furthermore, a similar (although non-significant) trend is seen for localization and externalization.

In Listening Test I of his doctoral thesis, Brinkmann demonstrates the advantage of paired BRIR and HCF sets [7]. In our study, there are two paired conditions (P); one personalized (C02) and one generic (C06). Similarly, there are two mismatched conditions (M); one with an individual BRIR (C03) and one with a generic BRIR (C05). For all three qualities, we see the following trends:

C02(P) > C03(M), C06(P) > C05(M)

Our results corroborate Brinkmann's finding that a paired generic set (C06) trends towards better performance than a mismatched set with generic BRIR and individual HCF (C05).

Our data demonstrates the perceptual advantage of a fully individualized binauralization process. C02 receives the highest MOS for all qualities and is the only condition to receive a MOS above 4.0 for all qualities. Furthermore, C02 is the only condition with statistically significant improvement over C04 for all qualities. There is, however, still room for improved personalization, particularly regarding coloration.

In conclusion, our results provide several rules of thumb for effective binaural reproduction. First, binauralization with a generic BRIR and no headphone compensation is relatively ineffective compared to the other explored methods. The inclusion of either individual or generic headphone compensation tends to improve binaural reproduction no matter the BRIR source. Furthermore, if a non-individual BRIR is used, a paired HCF measured on the same generic source is recommended. Finally, the combination of individualized BRIR and headphone compensation provides the most accurate binaural experience with respect to the discussed qualities; localization, externalization, and coloration.

7. FUTURE WORK

The listening test results presented in this study are specific to the measurement and inverse filter design methods defined in Sections 2 and 3, the Sennheiser HD800 headphones used for binaural reproduction, and the MLS test signal. Altering one or more of these may improve or diminish the reported localization, externalization, and/or coloration MOS.

Our data suggests that headphone compensation has a substantial impact on the coloration of a binaural signal. However, further examination of headphone compensation methods may reveal a perceptually superior HPTF measurement or inversion procedure. Section 2.2.1 introduced the concept of intra-individual HPTF variability. Additional investigation is required to develop an accurate method of quantifying and/or mitigating such variability. This may result in more accurate HCFs with precise equalization of high frequency pinna related notches.

Adding head-tracking to the assessment defined in Section 4 may reduce front-back confusion and improve the localization accuracy of binaural test signals. Furthermore, the development of a task-based evaluation platform (e.g. VR localization game as in [16]) will provide quantifiable measures of the benefits of individualization on a relevant use case.

Although we have concluded that an individualized binaural experience is preferred, a scalable and commercially viable system for the personalization of BRIRs and HCFs is yet to be developed. In future

work, we will use the methods and results of this study to assess new procedures for the measurement, computation, and selection of personalized BRIRs and HCFs. We hope that this research advances the development of transparent, accurate, and accessible systems for individualized binaural synthesis.

References

[1] Xie, B. (2013). *Head-Related Transfer Function and Virtual Auditory Display*. Plantation, FL: J. Ross Publishing.

[2] Møller, H. (1992). Fundamentals of Binaural Technology. *Applied Acoustics, 36*, 171-218. doi: 10.1016/0003-682X(92)90046-U.

[3] Zotkin, D.Y.N., Hwang, J., Duraiswaini, R. and Davis, L.S. (2003). HRTF personalization using anthropometric measurements. *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 157-160. doi: 10.1109/ASPAA.2003.1285855.

[4] Rothbucher, M., Veprek, K., Paukner, P., Habigt, T. and Diepold, K. (2013). Comparison of head-related impulse response measurement approaches. *The Journal of the Acoustical Society of America, 134*, 223-224. doi: 10.1121/1.4813592.

[5] Møller, H., Sørensen, M.F., Jensen, C.B. and Hammershøi, D. (1996). Binaural Technique: Do We Need Individual Recordings? *Journal of the Audio Engineering Society, 44(6)*, 451-469. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=7897.

[6] Schärer, Z. and Lindau, A. (2009). Evaluation of Equalization Methods for Binaural Signals, presented at the 126th AES Convention. Munich, Germany.

[7] Brinkmann, F. (2011). *Individual Headphone Compensation for Binaural Synthesis* (Doctoral dissertation). Retrieved from https://www2.ak.tu-berlin.de/~akgroup/.

[8] Farina, A. (2000). Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique, presented at the 108th AES Convention. Paris, France.

[9] Desloge, Joseph. "Pa-Wavplay." https://github.com/jgdsens/pa-wavplay.

[10] Brinkmann, F. and Weinzierl, S. (2017). AKtools – An Open Software Toolbox for Signal Acquisition, Processing, and Inspection in Acoustics, presented at the 142nd AES Convention. Berlin, Germany, 2017, e-Brief 309.

[11] International Telecommunication Union. (2015). Methods for the subjective assessment of small impairments in audio systems (ITU-R BS.1116-3). Geneva, Switzerland. Retrieved from https://www.itu.int/rec/R-REC-BS.1116-3-201502-I/en.

[12] Kuklarni, A. and Colburn, H.S. (2000). Variability in the characterization of the headphone transfer-function. *Journal of the Acoustical Society of America, 107(2)*, 1071-1074. doi: 10.1121/1.428571.

[13] Møller, H., Hammershøi, D. , Jensen, C.B. and Sørensen, M.F., (1995). Transfer Characteristics of Headphones Measured on Human Ears. *Journal of the Audio Engineering Society, 43(4)*, 203-217. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=7954.

[14] Norcross, S.G., Bouchard M., and Soulodre, G.A. (2006). Inverse Filtering design using a minimal phase target function from regularization, presented at the 121st AES Convention. San Francisco, CA.

[15] Tukey, J. (1949). Comparing Individual Means in the Analysis of Variance. *Biometrics, 5(2)*, 99-114. Retrieved from http://www.jstor.org/stable/3001913.

[16] Poirier-Quinot, D. and Katz, B. F.G. (2018). Impact of HRTF individualization on player performance in a VR shooter game I, presented at the Conference on Spatial Reproduction. Tokyo, Japan.