Audio Engineering Society

Conference Paper

Presented at the Conference on
Immersive and Interactive Audio
2019 March 27 – 29, York, UK

The effect of generic headphone compensation on binaural renderings

Isaac Engel, David Lou Alon, Philip W. Robinson, Ravish Mehra

Oculus & Facebook
1 Hacker Way, Menlo Park, CA 94025, USA

Correspondence should be addressed to David Lou Alon (davidalon@fb.com)

ABSTRACT

Binaural rendering allows us to reproduce auditory scenes through headphones while preserving spatial cues. The best results are achieved if the headphone effect is compensated with an individualized filter, which depends on the headphone transfer function, ear morphology and fitting. However, due to the high complexity of remeasuring a new filter every time the user repositions the headphone, generic compensation may be of interest. In this study, the effects of generic headphone equalization in binaural rendering are evaluated objectively and subjectively, with respect to unequalized and individually-equalized cases. Results show that generic headphone equalization yields perceptual benefits similar to individual equalization for non-individual binaural renderings, and it increases overall quality, reduces coloration, and improves distance perception compared to unequalized renderings.

1 Introduction

Binaural rendering allows us to reproduce auditory scenes through headphones while preserving all their spatial cues, by using measured or simulated binaural impulse responses [1, 2, 3], and is therefore widely used for synthesizing 3D sound in virtual and augmented reality applications. For instance, for a static listener and a single sound source in a room, the transfer function between the source and the listener's ears can be measured as a pair of filters (left and right), which are referred as the Binaural Room Impulse Response (BRIR) [2].

If a dry audio signal is convolved with a BRIR and presented through headphones, the listener should get the sensation that a real source is producing the sound at the corresponding location and with all the room acoustics preserved. However, for the synthesized signal to be indistinguishable from the real sound field, it should not be altered by the headphones, which have a non-flat frequency response. Therefore, it is necessary to use an equalization filter to compensate the effect of the headphone transfer function (HpTF). Essentially, the transfer function of the filter should be the inverse of the HpTF, so that when playing the equalized signal through the headphones, both transfer functions cancel out and the listener receives an unaltered version of the rendered binaural audio [4].

It is important to note that the HpTF compensation approach differs from other equalization methods recommended in the literature. A well known example is the Harman target curve, which was designed to simulate the response of a stereo loudspeaker system in a reverberant room, based on the assumption that music recordings are often optimized for such a setup [5]. In the case of binaural rendering, however, the room response is already included in the BRIR and therefore should not be taken into account when designing the headphone equalization curve.

The BRIR and HpTF are highly dependent on the morphology of the ear [6]; therefore, the highest degree of authenticity is only achieved when an individualized pair of BRIR and headphone equalization filter (HpEQ) are used [4]. Previous research has shown that when individualized filters are used static listeners could not distinguish between a real and a rendered audio source in a discrimination task [7]. Other studies have claimed that discrimination rates are higher (a) for broadband noise stimuli than for speech or music, (b) if listeners are given unlimited listening time, or (c) if head movements are allowed [8, 9].

However, if a non-individualized BRIR is used, it is no longer obvious which kind of HpEQ optimizes the quality of the binaural simulation. Lindau and Brinkmann [10] claim that the best practice is to use an HpEQ filter measured on the same head as the BRIR; the second-best choice would be an individualized HpEQ, and the least preferable option would be to use a non-individualized HpEQ measured on a different subject than the BRIR was. Nevertheless, in the perceptual evaluation, they used a single universal attribute to measure similarity between simulated and real audio, and it is therefore unclear whether listeners were paying more attention to the timbral characteristics of the audio content or to its spatial features. Also, in the ABC/HR type of comparison which was used, tested conditions were compared to the real loudspeaker but not to each other, so small perceptual differences between HpEQ types could have been lost.

In this study, the effects of individual and generic headphone compensation on the quality of a binaural rendering were further investigated, both objectively and perceptually. The contributions can be summarized as follows:

1. Instead of a single global rating, several features (overall similarity, coloration, distance, direction) were assessed separately, to better understand how all the binaural audio content characteristics are affected by the HpEQ type.
2. Evaluations were performed in both ideal and non-ideal binaural rendering scenarios (individualized and non-individualized BRIR, respectively).
3. A multiple stimuli test with hidden reference and anchor (MUSHRA), which is robust for measuring small and intermediate differences [11], was used. Furthermore, this test allowed for direct comparison between the different test conditions (individual equalization, generic equalization and unequalized).

2 Methods

A total of 12 subjects (ages 26-55, 2 female) took part in this study. All reported normal hearing and had previous experience with listening tests. Individual HpEQ filters and BRIRs were measured for all subjects with a pair of open headphones. Several test conditions were defined combining different types (generic/individual) of HpEQ and BRIR, which were analyzed objectively and subjectively in a perceptual experiment.

2.1 Hardware setup

Custom "floating" headphones were built, similar to the ones used by Langendijk and Bronkhorst [7] and by Romigh et al. [12], consisting of a pair of earbuds (Sennheiser MX475) attached to a headband (see Fig. 1). The purpose of this design was to leave the listener's ear canal unoccluded while minimizing the effect of the hardware on the head related transfer function (HRTF) [7]. Another reason for choosing these headphones

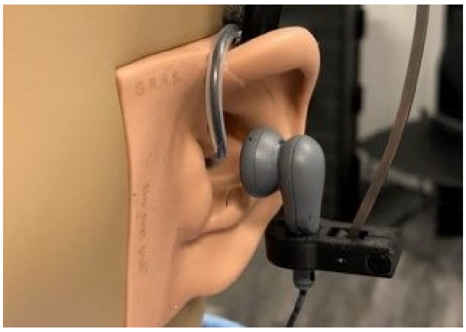


Fig. 1: Custom "floating" headphones (Sennheiser MX475) and binaural microphones (Brüel and Kjær 4101-B) mounted on KEMAR head and torso simulator (GRAS).

was that they have similar frequency response characteristics to air-conducted built-in speakers, which are prevalent in current virtual and augmented reality headsets such as the Oculus Go™, Microsoft HoloLens™, and Bose AR™. This consideration seemed relevant given that virtual and augmented reality are a common application area for binaural rendering.

Subjects were seated in the center of a reverberant room (RT30$_{[400Hz-1250Hz]}$ = 244 ms), wearing a pair of in-ear microphones (Brüel and Kjær 4101-B). A sound source (Genelec 8020) was placed at 45°azimuth, 0°elevation, and a distance of 2 meters from the subject's head. An adjustable chair and a laser alignment system were used to make sure that the head was at the right position and orientation during the measurements.

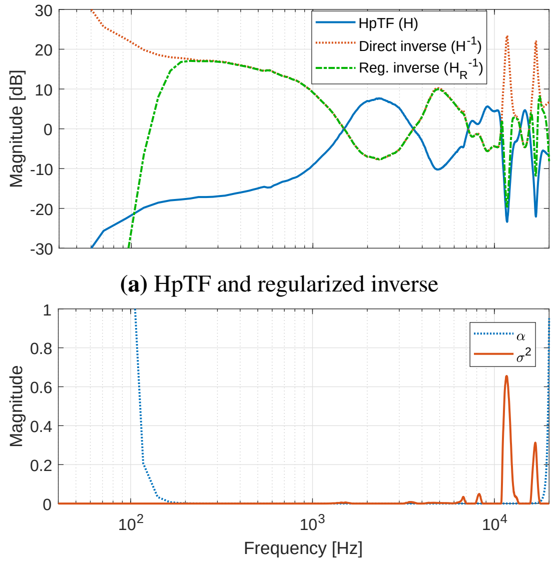
2.2 Headphone equalization approach

In order to compensate for the HpTF, headphone equalization (HpEQ) filters were calculated using frequency-dependent regularization [13], which has been shown to perform better than other methods in perceptual tests [14]. In general, the goal of regularization is to avoid the excessive boost of certain frequencies that happens if the direct inverse of the transfer function is used as a compensation filter, thus preventing distortion and sensitivity to measurement errors [13]. In the particular case of headphone compensation, regularization prevents the inversion of narrow notches at high frequencies, which could lead to ringing artifacts if the headphones are repositioned after the measurement [14]. For this reason, a frequency-dependent regularization parameter must be set to a higher value at frequencies where those notches are present. While this parameter has traditionally been adjusted by expert listeners [14], Bolaños et al. [15] have proposed a procedure by which to calculate it, demonstrating positive objective and perceptual results. In this study, the latter approach is used, calculating the regularized inverse $H_R^{-1}(\omega)$ of a headphone response $H(\omega)$ as

$$H_R^{-1}(\omega) = \frac{H^*(\omega)}{|H(\omega)|^2 + [\alpha(\omega) + \sigma^2(\omega)]} D(\omega) \qquad (1)$$

where $D(\omega)$ is a modeling delay to ensure that the filter is causal, $\alpha(\omega)$ is the parameter which defines the bandwidth and maximum amplification of the filter, and $\sigma^2(\omega)$ is an estimator of the amount of regularization needed within the inversion bandwidth. We define

$$\alpha(\omega) = \alpha_0 + \frac{1}{|W(\omega)|^2} - 1 \qquad (2)$$



(a) HpTF and regularized inverse

(b) Regularization parameters

Fig. 2: Example of HpEQ filter calculation. (a) Headphone transfer function (H), its direct inverse (H^{-1}) and regularized inverse (H_R^{-1}); (b) Regularization parameters α and σ^2 for that HpTF.

$$\sigma(\omega) = \begin{cases} |\hat{H}(\omega)| - |H(\omega)| & \text{if } |\hat{H}(\omega)| \geq |H(\omega)| \\ 0 & \text{if } |\hat{H}(\omega)| < |H(\omega)| \end{cases} \qquad (3)$$

$\alpha(\omega)$ is calculated from a unity-gain passband filter $W(\omega)$, which delimits the bandwidth within which the headphones are equalized. α_0 is a scalar that limits the amount of amplification allowed by the filter (0 means no limit). $\sigma(\omega)$ is defined as the negative deviation of the headphone response $H(\omega)$ from a smoothed version $\hat{H}(\omega)$, which will be larger in zones with narrow notches [15].

As seen in Fig. 2, the HpEQ filter (regularized inverse) is similar to the direct inverse of the HpTF, except that amplification is reduced outside the defined headphone bandwidth (200-20000 Hz in this case) and in zones with narrow notches; this is particularly noticeable around 11 and 17 kHz.

2.3 HpEQ and BRIR measurements

Measurements were performed with the sine sweep technique [16], using a sweep length of 2 s for the HpTF and 8 s for the BRIR, with a frequency range from 10 to 24000 Hz. HpEQ filters (H_R^{-1}) were calculated from the HpTF (H) following Eqs. 1-3, with the following parameters:

1. $W(\omega)$: 5th order Butterworth bandpass filter (200-20000 Hz), according to the headphone bandwidth.
2. α_0: $2.5 \cdot 10^{-4}$, which limits the amplification to 30 dB within the inversion range.
3. $\hat{H}(\omega)$: smoothing window of 1 ERB (equivalent rectangular bandwidth), following [17], which gave good results in preliminary tests.

The HpTF and BRIR were measured for each subject and for a KEMAR head and torso simulator (GRAS), which was used for the "generic" conditions. Measurements on subjects were done immediately before performing the test, and headphones were not repositioned or removed until the experiment was finished. An overview of the results is detailed in Section 3 and shown in Fig. 4.

2.4 Audio material

In order to evaluate the effect of HpEQ for various scenarios, three different dry audio materials were used in the listening test:

1. *Speech*: anechoic male speech recording[1].
2. *Pink noise*: a sequence of 4 broadband noise bursts of length 750 ms with 500 ms of silence between them. Each noise pulse was faded in and out with a 50 ms raised cosine window.
3. *Guitar*: anechoic guitar recording[1].

All signals were of length 5 s and were faded in and out with 50 ms raised cosine windows. To ensure that all spectral content could be played at a reasonable level (60 dBA) without distortion, a band-pass filter between 500 and 16000 Hz was applied (see Fig. 3).

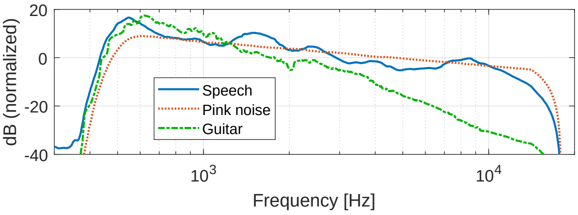


Fig. 3: Spectra of the audio materials. Curves were normalized and smoothed with a third-octave window.

2.5 HpEQ and BRIR test conditions

In this study, we wanted to evaluate the effect of HpEQ for the ideal case where the individualized BRIR is available, as well as for the case where a generic BRIR is used. Therefore, two independent variables were tested:

1. BRIR type: individualized (IndBRIR) / generic (GenBRIR).
2. HpEQ type: individualized (IndHpEQ) / generic (GenHpEQ) / no (NoHpEQ).

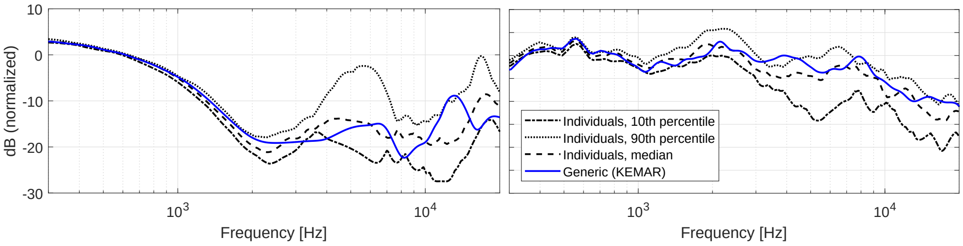
This makes a total of six test conditions, from now on referred to as xxxBRIR+yyyHpEQ (e.g., IndBRIR+GenHpEQ).

Although the study initially included the real loudspeaker as a reference condition, it was finally removed for two reasons: (1) for a fair comparison of the source direction between the loudspeaker and rendering, the listener's head needed to be completely static (i.e., with the help of a chin rest) during the whole duration of the experiment, which was found to be sufficiently uncomfortable so as to introduce a bias in the responses, and (2) IndBRIR+IndHpEQ was considered a suitable reference condition, as it was found to be almost indistinguishable from the real loudspeaker in a preliminary study, in agreement with findings from Langendijk and Bronkhorst [7].

3 Objective evaluation

title Spectra of the three dry audio materials are shown in Fig. 3. It can be seen that *pink noise* and *speech* offer the widest content in terms of frequency range, and therefore may be more effective in revealing coloration changes across different test conditions.

[1] pcfarina.eng.unipr.it/Public/Aurora_CD/Anecoic/Archimedes/CD-cover/Archimedes.htm



(a) HpEQ **(b)** BRTF

Fig. 4: Statistics of (a) calculated HpEQ filters and (b) Binaural Room Transfer Function (BRTF), for the left ear. Median, 10th and 90th percentiles across subjects are indicated with dashed lines; generic is indicated with a solid line. Curves were normalized and smoothed with a third-octave window.

The *guitar* stimulus, on the other hand, might give the listeners less spectral information, which could make the perceptual evaluation more challenging.

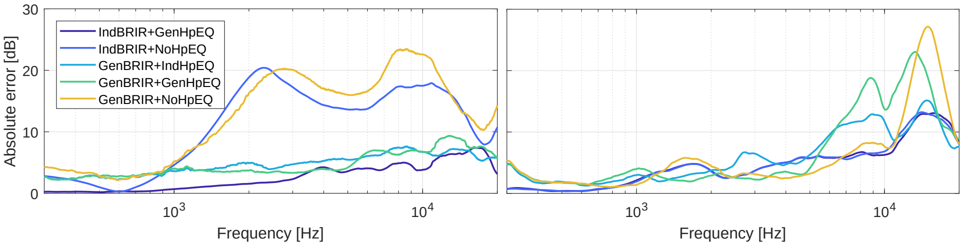
Figure 4 shows the statistics (10th and 90th percentiles and median) of HpEQ filters (Fig. 4a) and the BRIRs in the frequency domain or Binaural Room Transfer Functions (BRTF, Fig. 4b), based on measurements from the 12 subjects. In addition, the generic HpEQ filter and BRTF measured on KEMAR are also shown. In Fig. 4a, it can be observed how curves are very similar across subjects for frequencies below 2 kHz, and start to diverge above that point, which indicates that individual features tend to appear at higher frequencies, as found by Pralong and Carlile [4]. The overall shape of the HpEQ curves hints at the limitations of the custom “floating” headphones when unequalized, given the >20 dB difference in gain between the low (300 Hz) and mid (2 kHz) frequencies.

In Fig. 4b, similarly to the HpEQ plot, curves seem to converge in the lower part of the spectrum, and diverge for higher frequencies. This is due to the nature of the HRTFs, which have small variability at low frequencies and higher variability at high frequencies. It can be seen that the generic BRIR differs from the individual measurements by up to 5 dB for frequencies as low as 1.5 kHz, and by up to 15 dB for higher frequencies.

To get a better understanding of the impact of the HpEQ and BRIR on the in-ear pressure level, and perhaps some insights on the perceived differences between the tested conditions, it may be interesting to analyze their effect on monaural and binaural cues. To that end, the “complete” binaural response – taking into account both the BRIR and the HpEQ filter – was computed for each test condition, across all subjects. Taking the IndBRIR+IndHpEQ response as a reference, an error metric can be calculated as the spectral difference between the log magnitude response of the reference and the log magnitude response of each of the tested conditions.

Figure 5a shows the absolute spectral difference for (a) the complete binaural response and (b) the interaural level difference (ILD) [18] (Eq. 1.12). Each curve was constructed by taking the median spectral difference across all subjects for each frequency bin, and smoothing the result with a third-octave window. As anticipated, NoHpEQ conditions showed the largest deviations from the reference, with errors of the order of 20 dB, particularly above 2 kHz. Taking into account the variability of the results across individuals, and the error introduced by the generic BRIR, the differences between IndHpEQ and GenHpEQ are relatively small. The performance of each tested condition on the subjective evaluation may be predicted by observing these results – e.g., large errors may be perceived by listeners as severe coloration changes.

A similar spectral difference error metric can be calculated for the ILD, by subtracting the ILD of each tested condition from the reference IndBRIR+IndHpEQ. Given that filters were independently designed for left and right ears, it was fair to assume that ILDs changed from one condition to another. Figure 5b shows the absolute ILD error of each condition, taking IndBRIR+IndHpEQ as a reference. A higher error was observed above 6 kHz for GenBRIR than for IndBRIR conditions, with GenBRIR+GenHpEQ and GenBRIR+NoHpEQ having the largest error (above 11 kHz). This result can be explained by the HRTF variability across different subjects, which is generally dominant at higher frequencies.



(a) Median absolute monaural error **(b)** Median absolute ILD error

Fig. 5: Error curves for all test conditions: (a) absolute monaural error; (b) absolute ILD error. Curves were calculated by convolving the BRIR and HpEQ filter, both dependent on the test condition, then subtracting the reference curve (IndBRIR+IndHpEQ), and finally taking the median value across subjects for each frequency bin (final result smoothed with a third-octave window).

Interaural time difference (ITD) analysis was considered as well, following the recommendations by Katz and Noisternig [19]. However, impulse response onset detection was found to be problematic on the contralateral ear, perhaps due to the non-anechoic conditions of the measurements not providing a high enough signal-to-noise ratio. It was therefore decided to leave ITD analysis for future follow-up studies.

4 Subjective evaluation

4.1 Method

A MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test, as defined in ITU-R BS.1534-3 [11], was used to perform the perceptual evaluation. The MUSHRA paradigm was chosen because it allows the listener to compare all the test conditions to one another, in addition to the reference, which makes it easier to detect small and intermediate differences between them.

Three different dry audio signals and four perceptual attributes where assessed, making a total of 12 trials per participant. Each subject performed the experiment in a single session, which lasted approximately 45 minutes and included around 5 minutes for training and 40 minutes for evaluation and ranking.

4.1.1 Assessed attributes

A preliminary listening test was run, where subjects were asked to point out and rate the most relevant attributes across the different listening conditions (individual HpEQ, generic HpEQ, no HpEQ). To keep the length of the experiment reasonable, the four most dominant attributes were chosen, which are the following:

1. Overall similarity (OVS): Any and all detected differences between the reference and the tested signal.

2. Coloration (COL): Any and all detected differences in timbral impression and tonalness between the reference and the tested signal.

3. Distance (DIS): Whether the tested signal is perceived at the same distance as the reference, regardless of the direction of incidence.

4. Direction (DIR): Whether the tested signal's angle of incidence (azimuth/elevation) is the same as for the reference.

This attribute choice is similar to the findings of Brinkmann et al. [9], who found that the most relevant attributes when rating binaural renderings were *difference*, *high frequency color*, *brightness*, *pitch*, and *distance*. It was also similar to the ITU-R BS.1534-3 [11] recommendation, which suggests using *basic audio quality*, *timbral quality*, *localization quality* and *environment quality*. Participants were provided an informative sheet with attribute definitions and examples, extracted from the Spatial Audio Quality Inventory (SAQI) by Lindau et al. [20].

4.1.2 Training stage

Participants performed a training phase, where they were exposed to all the signals which they would later experience during the test. A screen was presented, where all the combinations of audio material and test conditions were available as buttons, and could be played as many times as needed by clicking them. Buttons were unlabeled and their order was randomized to avoid introducing bias. Participants were encouraged to spend as long as they needed to familiarize themselves with the test material.

4.1.3 Evaluation stage

After the training, participants proceeded to the evaluation stage. This consisted of 12 trials, one for each combination of audio material and attribute. The evaluation was divided into four blocks, one for each attribute. The first block was always OVS, while the other three were presented in a random order. Within each block, the order of the trials (one per audio material) was also randomized. In each trial, subjects were presented an audio material and could switch at will between all the test conditions (BRIR/HpEQ types). A reference was provided, which was always the IndBRIR+IndHpEQ condition, and subjects were asked to rate the similarity of each condition against the said reference, according to the current attribute (OVS/COL/DIS/DIR), with a score from 0 to 100. A rating of 100 would mean that the signal and the reference were identical for that particular attribute.

Subjects were informed that one of the signals was a hidden reference, and therefore they were required to give at least one rating of 100. All signals could be played as many times as needed. It was possible to switch between signals during the playback; a 2 ms raised cosine fade in/out with no crossfade was applied in the transition between audio signals. A practice trial, which did not count towards the results, was presented at the beginning of the evaluation stage so the subjects could familiarize themselves with the rating procedure.

Although in MUSHRA tests it is recommended to use at least one low-passed version of the reference signal as a "low-quality" anchor [11], in the proposed experiment it was not trivial to define a proper anchor for attributes such as DIR or DIS. For instance, a low-passed version of the reference (IndBRIR+IndHpEQ) would still preserve the binaural and monaural cues of the individualized BRIR, so the perceived direction of the sound source would remain unaffected. Therefore, it was decided not to use any anchor in this study.

4.2 Results

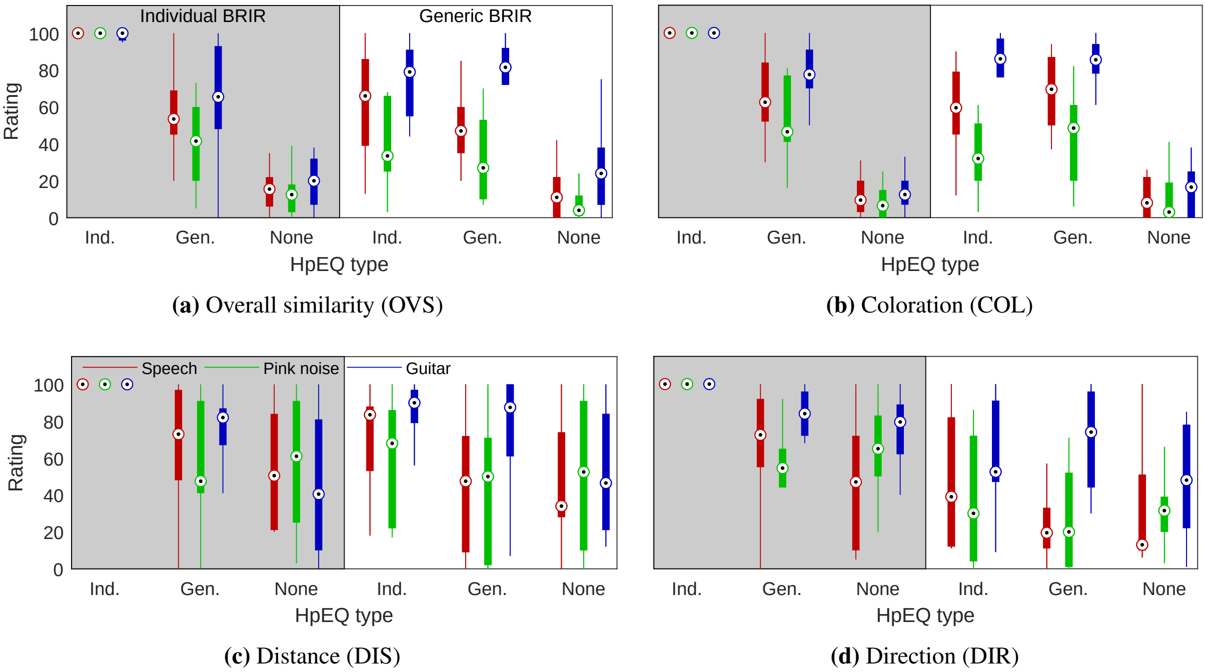
Figure 6 provides an overview of the subjective evaluation results across 10 subjects, divided by test condition and audio material. (two individuals were excluded in post-screening because they failed to rate the reference above 90 points for more than 15% of the trials [11]). Interquartile ranges are indicated with boxes, medians with circles, and the most extreme data points with thin lines. BRIR and audio material are denoted by background and line colors, respectively. A quick inspection reveals that IndBRIR+IndHpEQ was correctly identified as the hidden reference condition by most listeners, as it was rated consistently close to 100. Data for other conditions shows considerable spread, which makes it difficult to extract conclusions from the boxplot alone. This spread is caused by the individual bias caused by each subject's own rating criteria. Thus, the appropriate way to analyze the data is through a statistical method for dependent samples, where all ratings by the same subject are considered as dependent on each other – which in this case would be a repeated measures analysis of variance (rmANOVA) [11].

A three-way rmANOVA was run for each of the four attributes (OVS/COL/DIR/DIS), where HpEQ, BRIR and audio material were the between-subject factors. A significance level of 0.05 was used.

4.2.1 Effect of BRIR

A significant effect of BRIR type was found on all attributes $[F(1,9) > 11.91, p < 0.008]$, as well as a strong interaction between BRIR and HpEQ type $[F(2,18) > 5.58, p < 0.02]$. This interaction was largely caused by the IndBRIR+IndHpEQ (reference) condition obtaining very high and consistent ratings, therefore biasing the results of the ANOVA.

When considering only GenHpEQ and NoHpEQ data, the effect of the BRIR was found to be significant only on DIR $[F(1,9) = 14.49, p = 0.004]$. Post-hoc dependent samples t-tests show that for this attribute, individual BRIRs got higher ratings than the generic BRIR, independently of the audio material and HpEQ type. No significant effect of the BRIR was found on OVS, COL or DIS in the same analysis.



(a) Overall similarity (OVS)

(b) Coloration (COL)

(c) Distance (DIS)

(d) Direction (DIR)

Fig. 6: Results of the MUSHRA test for each of the four assessed attributes. Interquartile ranges are indicated with boxes, medians with circles, and the most extreme data points with thin lines. Background color (grey/white) denotes the BRIR type. Audio materials are, from left to right, *speech* (red), *pink noise* (green) and *guitar* (blue).

4.2.2 Effect of HpEQ

A significant effect of the HpEQ type was found for all attributes [$F(2,18) > 6.05, p < 0.01$]. As mentioned above, the significant interaction between HpEQ and BRIR made it necessary to perform separate analyses for the different conditions.

The difference between IndHpEQ and GenHpEQ was found to be significant on IndBRIR conditions for all attributes and audio materials [$F(1,9) > 15.23, p < 0.004$], and non significant on GenBRIR conditions.

On the other hand, GenHpEQ ratings were found to be significantly higher than NoHpEQ ones on OVS and COL [$F(1,9) > 30.69, p < 0.001$], but not on DIS or DIR. Dependent samples t-tests showed that on OVS and COL the difference was significant for all audio materials and both BRIR types [$t(19) > 4.82, p < 0.001$]. Further, the effect on DIS, the effect was found to be significant for *guitar* [$t(19) = 3.36, p = 0.003$] and not significant for *speech* or *pink noise*.

5 Discussion

Results of the subjective evaluation indicate that BRIR type was the dominant factor in perceived direction, which agrees with the common consensus that an individualized HRTF should always yield more accurate sound localization than a generic one [21]. This result is commensurate with the observations made in the objective evaluation, where GenBRIR conditions were found to have a higher ILD error than IndBRIR ones, which probably caused a larger error in the perceived source's azimuth.

Another finding was that coloration was mainly affected by the HpEQ type, and the largest differences are perceived in the NoHpEQ condition. This result could also have been anticipated from the objective evaluation, given that the largest monaural errors were observed on unequalized conditions.

The differences between IndHpEQ and GenHpEQ seem to be very evident when using the individual BRIR, which can be explained by the fact that the reference is mostly rated 100, while the other condition has the subject's individual bias within it, but are less accentuated when using the generic BRIR. Actually, results suggest that listeners did not perceive any improvement when using their own HpEQ filter on generic-BRIR conditions, not only in terms of overall similarity to a reference (which was previously shown by Lindau and Brinkmann [10]) but also for specific attributes such as coloration, distance and direction. The latter two are particularly interesting because they imply that the spatial perception of a non-individualized rendering is not significantly altered if a generic HpEQ is used instead of an individualized one.

It is noteworthy that the correlation between OVS and COL is higher (Pearson correlation coefficient $\rho = 0.76$) than the correlation between OVS and DIR ($\rho = 0.52$) or OVS and DIS ($\rho = 0.48$), which may indicate that, when rating overall similarity, subjects were paying more attention to coloration changes than to spatial features. This would be supported by the results from Brinkmann et al. [9], where coloration and timbre-related attributes were the ones given the most weight by listeners when comparing binaural content.

Interestingly, DIS, which was chosen as an intuitive attribute to evaluate externalization, is partially affected by the HpEQ type, as GenHpEQ obtained higher ratings than NoHpEQ for some audio materials. This trend may indicate that subjects externalized the equalized audio content better than the non-equalized one. Several subjects claimed in informal discussions after the experiment that it was harder for them to externalize the broadband noise than the other audio materials. Such statements would support the observed trend, given that *pink noise* stimuli received lower DIS ratings than *speech* or *guitar* ones.

Finally, the lack of evidence of any effect of HpEQ on DIR indicates that (a) the HpEQ filter did not alter the interaural cues enough to be perceivable, which is supported by the objective evaluation, where ILD error was not found to be significantly affected by HpEQ type, and (b) perception of elevation did not change even though monaural cues were altered by the HpEQ filter. This might be partially explained by the fact that the loudspeaker was visible to the subjects at all times, so it is possible that a "ventriloquist effect" [22] was taking place, "pulling" the sounds towards the loudspeaker as the most plausible visible source. It would have been interesting to test more source directions. However, due to the relatively high number of independent variables, testing more directions is suggested for future experiments. One possible follow-up to this study would be to evaluate the same set of HpEQ types in a sound localization task, which would provide a better understanding of the impact of a fixed headphone compensation filter on elevation perception for a static listener.

Overall, results suggest that generic headphone equalization offers perceptual benefits over unequalized content in terms of overall quality, coloration and, for certain audio contents, perceived distance. Thus, generic equalization may have a positive effect, both in the perceived quality and in the accuracy of the spatial features of binaural content, and therefore it might be beneficial to use:

1. When presenting a non-individualized binaural rendering,

2. When presenting an individualized binaural rendering and an individual HpEQ is not available.

6 Summary

In this study, the effect of generic headphone compensation on binaural renderings was explored. Individual binaural room impulse respones (BRIRs) and headphone transfer functions (HpTFs) were measured for several listeners and a "generic" dummy head. Then, an objective analysis and a perceptual evaluation were performed to compare different combinations of generic and individual BRIRs and headphone equalization (HpEQ) filters.

Results show that, although the best quality was achieved for the case of binaural rendering with individualized BRIR and HpEQ filters, generic HpEQ still yielded significant perceptual benefits compared to unequalized reproduction, including a reduction in coloration, an increase in overall quality, and an improved perception of distance. Also, it was found that for non-individualized renderings generic equalization provided benefits at a level that was similar to individualized equalization for all the tested attributes (overall similarity, coloration, distance and direction).

References

[1] Wightman, F. L. and Kistler, D. J., "Headphone simulation of free-field listening. I: stimulus synthesis," *The Journal of the Acoustical Society of America*, 85(2), pp. 858–867, 1989.

[2] Møller, H., "Fundamentals of binaural technology," *Applied acoustics*, 36(3-4), pp. 171–218, 1992.

[3] Begault, D. R. and Trejo, L. J., "3-D sound for virtual reality and multimedia," 2000.

[4] Pralong, D. and Carlile, S., "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space," *The Journal of the Acoustical Society of America*, 100(6), pp. 3785–3793, 1996.

[5] Olive, S., Welti, T., and McMullin, E., "Listener preferences for different headphone target response curves," in *Audio Engineering Society Convention 134*, Audio Engineering Society, 2013.

[6] Møller, H., Hammershøi, D., Jensen, C. B., and Sørensen, M. F., "Transfer characteristics of headphones measured on human ears," *Journal of the Audio Engineering Society*, 43(4), pp. 203–217, 1995.

[7] Langendijk, E. H. A. and Bronkhorst, A. W., "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *The Journal of the Acoustical Society of America*, 107(1), pp. 528–537, 2000.

[8] Oberem, J., Masiero, B., and Fels, J., "Experiments on authenticity and plausibility of binaural reproduction via headphones employing different recording methods," *Applied Acoustics*, 114, pp. 71–78, 2016.

[9] Brinkmann, F., Lindau, A., and Weinzierl, S., "On the authenticity of individual dynamic binaural synthesis," *The Journal of the Acoustical Society of America*, 142(4), pp. 1784–1795, 2017.

[10] Lindau, A. and Brinkmann, F., "Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings," *The Journal of the Acoustical Society of America*, 60(1), pp. 54–62, 2012.

[11] "ITU-R BS. 1534-3: Method for the subjective assessment of intermediate quality level of audio systems," 2015.

[12] Romigh, G. D., Brungart, D. S., and Simpson, B. D., "Free-Field Localization Performance With a Head-Tracked Virtual Auditory Display," *IEEE Journal of Selected Topics in Signal Processing*, 9(5), pp. 943–954, 2015.

[13] Kirkeby, O. and Nelson, P. A., "Digital Filter Design for Inversion Problems in Sound Reproduction," *Journal of the Audio Engineering Society*, 47(7/8), pp. 583–595, 1999.

[14] Schärer, Z. and Lindau, A., "Evaluation of Equalization Methods for Binaural Signals," in *AES 126th Convention*, 2009.

[15] Bolaños, J. G., Mäkivirta, A., and Pulkki, V., "Automatic regularization parameter for headphone transfer function inversion," *Journal of the Audio Engineering Society*, 64(10), pp. 752–761, 2016.

[16] Farina, A., "Advancements in impulse response measurements by sine sweeps," in *Audio Engineering Society Convention 122*, Audio Engineering Society, 2007.

[17] Breebaart, J. and Kohlrausch, A., "The perceptual (ir) relevance of HRTF magnitude and phase spectra," *PREPRINTS-AUDIO ENGINEERING SOCIETY*, 2001.

[18] Xie, B., *Head-related transfer function and virtual auditory display*, J. Ross Publishing, 2013.

[19] Katz, B. F. and Noisternig, M., "A comparative study of interaural time delay estimation methods," *The Journal of the Acoustical Society of America*, 135(6), pp. 3530–3540, 2014.

[20] Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., and Weinzierl, S., "A spatial audio quality inventory (SAQI)," *Acta Acustica united with Acustica*, 100(5), pp. 984–994, 2014.

[21] Møller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D., "Binaural technique: Do we need individual recordings?" *Journal of the Audio Engineering Society*, 44(6), pp. 451–469, 1996.

[22] Alais, D. and Burr, D., "The ventriloquist effect results from near-optimal bimodal integration," *Current biology*, 14(3), pp. 257–262, 2004.