OPEN ACCESS Freely available online

I. Engel, D. L. Alon, K. Scheumann, J. Crukley and R. Mehra, "On the Differences in Preferred Headphone Response for Spatial and Stereo Content"
J. Audio Eng. Soc., vol. 70, no. 4, pp. 271–283, (2022 April).
DOI: https://doi.org/10.17743/jaes.2022.0005

On the Differences in Preferred Headphone Response for Spatial and Stereo Content

ISAAC ENGEL, DAVID L. ALON,* *AES Member*, **KEVIN SCHEUMANN, JEFF CRUKLEY, AND**
(isakengel@gmail.com) (davidalon@fb.com) (kevinscheumann@fb.com) (jcrukley@gmail.com)

RAVISH MEHRA, *AES Associate Member*
(ravish.mehra@fb.com)

Reality Labs, Meta, 1 Hacker Way, Menlo Park, CA 94025, USA

When reproducing spatial audio over headphones, ensuring that these have a flat frequency response is important to produce an accurate rendering. However, previous studies suggest that, when reproducing nonspatial content such as stereo music, the headphone response should resemble that of a loudspeaker system in a listening room (e.g., the so-called Harman target). It is not yet clear whether a pair of headphones calibrated in such way would be preferred by listeners for spatial audio reproduction too. This study investigates how listeners' preference regarding headphone frequency response differs in the cases of stereo and spatial audio content reproduction, rendered using individual binaural room impulse responses. Three listening tests that evaluate seven different target headphone responses, two headphones, and two reproduction bandwidths are presented with over 20 listeners per test. Results suggest that a flat headphone response is preferred when listening to spatial audio content, whereas the Harman target was preferred for stereo content. This effect was found to be stronger when user-specific equalization was used and was not significantly affected by the choice of headphone or reproduction bandwidth.

0 INTRODUCTION

In audio content production, spatial impression can be achieved in different ways, depending on the content type. In the case of traditional, nonspatial audio content (e.g., stereo), this can be done through amplitude panning (for left–right distribution of phantom sources) and audio effects (e.g., adding reverberation to modify the distance of phantom sources). Alternatively, spatial audio achieves it by generating binaural signals, i.e., sound pressure evaluated at the listener's ears, which inherently contain spatial cues [1]. Binaural signals can by either measured acoustically via in-ear microphones or generated by filtering the audio content with head-related impulse responses (HRIRs), which replicate the effect of the listener's anatomy on the sound as it reaches the ears [1]. When the room response is also included in these filters, they are referred to as binaural room impulse responses (BRIR).

Depending on the audio content type and its corresponding spatialization strategy, its spectral properties are affected differently. On the one hand, nonspatial audio production generally relies on using audio equipment with a flat frequency response both on the recording (microphones) and reproduction stages (loudspeakers). At the same time, such audio content is generally optimized for calibrated loudspeakers in a listening room (e.g., see [2]), meaning that the optimal listening conditions include the spectral coloration introduced by the interaction between the speakers, the room, and the listener's anatomy. On the other hand, spatial audio content's frequency response is affected by HRIRs' spectra (also known as head-related transfer functions or HRTFs), with the most salient spectral feature being a strong peak in the region of 3 kHz [3], and it is generally intended to be reproduced via headphones.

Because the purpose of spatial audio reproduction is to reproduce the binaural signals with high fidelity, the potential effect of the headphones on the signal must be accounted for. The headphone transfer function (HpTF), which quantifies the linear transformation between the digital signal sent to the headphones and the binaural signal measured at the ears, plays an important role in the reconstruction of binaural and monoaural cues [1, 4]. For an accurate reproduction of binaural signals, the HpTF must be compensated by filtering the audio content with headphone equalization (HpEQ) filters in order to produce a neutral (*flat*) frequency response. Because the HpTF depends on the listener's pinnae morphology and headphone fitting, optimal

*Corresponding Author.

equalization is achieved when individual filters are used [4, 5]. However, generic HpEQ has also been shown to be beneficial by improving perceived quality, coloration, and externalization for open-ear headphones [6, 7].

Whether individually generated or not, a *flat* HpTF may not be the optimal choice for the reproduction of nonspatial audio content, such as stereo music [8]. Møller et al. [9] argued that in such cases, the target HpTF (or, simply, "target") should be the frequency response of the system formed by the loudspeakers and the listener's head in free or diffuse fields. A later study by Lorho [10] attempted to parametrize such a diffuse field target by means of a single peak filter and proposed a refined version based on listener preferences. More recently, Olive et al. [11] proposed the *Harman* target, which was based on acoustical measurements in a calibrated listening room and showed that it was preferred by listeners over the targets from Møller et al. [9] and Lorho [10] for the reproduction of stereo music content. Later, perceptual studies by Olive et al. showed that variations of the *Harman* target were generally preferred to three over-ear commercial headphones in a study with 238 listeners [12] and to 30 in-ear headphones in a study with 71 listeners [13] for stereo music reproduction.

From previous literature it seems, therefore, that the appropriate target depends on the audio content type: a *flat* target is recommended for spatial audio reproduction, whereas other alternatives, such as the *Harman* target, may be better suited for nonspatial audio. However, to the best of the authors' knowledge, the effect of audio content type on target preference has not been thoroughly studied as of yet. This is relevant for the calibration of devices intended to reproduce both spatial and nonspatial audio content, such as augmented reality (AR) and virtual reality (VR) headsets.

The goal of this study is to assess whether the preferred target HpTF varies depending on the audio content type (stereo or spatial). For this purpose, several targets were perceptually evaluated in three double-blind listening tests, in which listeners were asked to rate each target according to their preference for several excerpts of stereo and spatial audio content, under different conditions. Preliminary results were presented in [14], and showed a significant effect of the audio content type on the preferred target in the particular case of a VR headset and employing an individual HpEQ approach. Said study had some limitations, such as not investigating the potential effect of headphone type and reproduction bandwidth and the fact that the variance across the tested targets may have been too large to observe finer differences between them. In this study, additional variables are introduced, such as two headphone types, an alternative selection of target HpTFs, and two different reproduction bandwidths, all of which serves to provide further insight on the research question.

The rest of the paper is structured as follows: SEC. 1 describes the conditions under which listener preference was assessed, SEC. 2 presents the methods, including the experimental setup, acoustical measurements and test procedure; SEC. 3 and SEC. 4 describe the three listening tests along with a discussion of their results; and SEC. 5 summarizes the findings and concludes the paper.

Table 1. Test conditions used in the three listening tests (LT). LT 1: target (large variance) vs. content type. LT 2: target (small variance) vs. content type. LT 3: target (small variance) vs. bandwidth. EQ, equalization; Gen., generic; Ind., individual.

		LT 1	LT 2	LT 3
Headphones	Audeze			x
	Headset	x	x	
Target	Ind. flat	x		
	Gen. flat	x	x	x
	Harman	x	x	x
	−1 Harman		x	x
	1/2 Harman		x	x
	2 Harman	x	x	x
	No EQ	x		
Bandwidth	Full			x
	Limited	x	x	x
Content	Stereo	x	x	x
	Spatial	x	x	

1 TEST CONDITIONS

In this study, listener preference was assessed under various conditions, namely the following:

1) Two headphones: a custom VR headset prototype with built-in loudspeakers and a pair of high-end over-ear headphones.
2) Seven HpTF targets.
3) Two reproduction bandwidths: one covering the full audible spectrum and other matching the limitations of an open-ear VR headset.
4) Two audio content types: stereo and spatial.

These are summarized in Table 1 and will be described in detail in the following subsections.

1.1 Headphones

Two different binaural reproduction systems (hereafter referred to as "headphones" for simplicity, even though one of them is not strictly a pair of headphones) were selected for this study:

1) Custom headset: Custom VR headset prototype with open-ear built-in loudspeakers, which had the frontal part removed to let the user see through. Its built-in loudspeakers are integrated in the headband and sit above the pinnae, and were wired to an external amplifier. The device provided realistic audio reproduction bandwidth limitations typical of commercial VR headsets (e.g., see Fig. 3 bottom left), due to the loudspeakers' size and their off-ear location.
2) Audeze LCD-2: Pair of high-end, over-ear, planar magnetic headphones that were used in a previous study [11].

These devices were chosen to represent two different typical user scenarios. The custom headset had an open-ear design typical of current AR/VR devices, with its consequent limitations: a potentially high crosstalk between channels

Table 2. Audio material used in the listening tests.

Key	Artist	Track	Album	Description
JW	Jennifer Warnes	*Bird on a Wire*	*Famous Blue Raincoat*	Pop with female vocals
SD	Steely Dan	*Cousin Dupree*	*Two Against Nature*	Pop with male vocals
GA	Stu Phillips	*Main Theme*	*Battlestar Galactica OST*	Classical Orchestra
SP	Neil Thompson	*List 1*	*Harvard Sentence Lists*	Male speech

and a narrow reproduction bandwidth. On the other hand, the Audeze pair was chosen to represent traditional over-ear headphones, which typically have a larger bandwidth and dynamic range, as well as lower distortion and crosstalk, than open-ear devices.

1.2 Targets

Based on the literature review presented here and for the purpose of this study, it is assumed that Harman and flat are the optimal targets for the reproduction of stereo and spatial content, respectively. Thus, it is hypothesized that listeners' preference ratings will be highest for one of these two targets, depending on the audio content type. It is also hypothesized that the preference rating of a given target will decrease in proportion to its deviation from the optimal target, which will vary depending on the audio content type. Note that this second hypothesis held true in a previous study, but it only considered stereo content and in-ear headphones [13]. To test the two hypotheses, several targets were constructed as linear combinations of Harman and flat as follows:

$$T(\omega, \gamma) = F(\omega) + \gamma\, Ha(\omega), \tag{1}$$

where ω is the frequency, $F(\omega)$ is the log-magnitude of the flat target (0 dB for all frequencies), $Ha(\omega)$ is the log-magnitude of the *Harman* target (from [15]), γ is a scalar, and $T(\omega, \gamma)$ is the resulting target. Therefore, when γ approaches zero, T will be closer to *flat*, while when γ approaches one, T will be more similar to *Harman*. On this basis, the following targets were evaluated:

1) *Flat* [$T(\omega, 0)$]: expected to be preferred for spatial content.
2) *Harman* [$T(\omega, 1)$]: expected to be preferred for stereo content.
3) *−1 Harman* [$T(\omega, -1)$]: closer to *flat* than to *Harman*.
4) *1/2 Harman* [$T(\omega, 0.5)$]: "half way" between *flat* and *Harman*.
5) *2 Harman* [$T(\omega, 2)$]: closer to *Harman* than to *flat*.
6) *No EQ*: measured magnitude response of the unequalized custom headset, chosen as a low-quality anchor condition.

Two versions of the *flat* target were implemented, bringing the total number of targets to seven. The first was generic (*gen. flat*), for which HpEQ filters were generated from a non-individual measurement (see [7]), and the second was individual (*ind. flat*), for which HpEQ filters were generated for the individual listeners. The *ind. flat* target has been shown to be optimal for the reproduction of spatial audio content [4] and was included in order to draw comparisons to previous studies. All other targets (*Harman*, *−1 Harman*, *1/2 Harman*, *2 Harman*, and *No EQ*) were generated from generic (i.e., nonindividual) measurements, as explained in more detail in SEC. 2.1.

1.3 Reproduction Bandwidth

Targets were defined for two different reproduction bandwidths: (i) *full*, which covered the full audible spectrum from 20 Hz to 24 kHz, and (ii) *limited*, which matched the bandwidth limitation of the custom headset's built-in loudspeakers by applying a third-order band-pass filter from 120 to 11,300 Hz. Fig. 1 shows the frequency responses of all the evaluated targets for full and limited bandwidths.

1.4 The Ear Canal Reference Point

A comparison between different targets requires a clear and consistent definition of how the measurement is performed. The *Harman* target is defined for headphones measured at the eardrum reference point (DRP) of a dummy head with ear canal simulators [11]. On the other hand, in

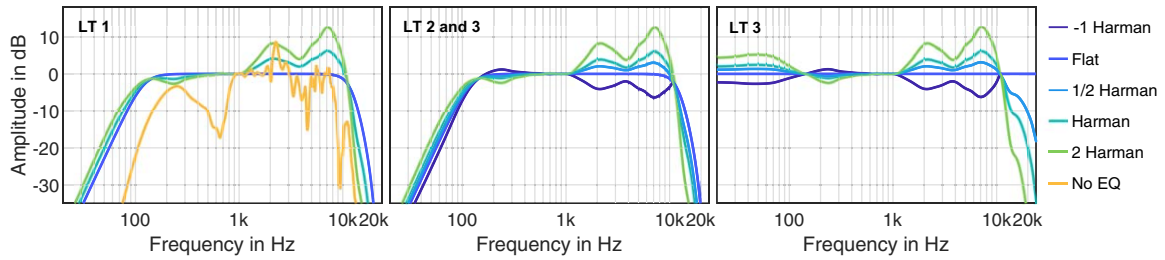

Fig. 1. Frequency responses of targets that were evaluated in the three listening tests (LT), all specified at the ear canal reference point (ECRP) (see SEC. 1.4). Left: in LT 1. Middle: in LT 2 and 3 (limited bandwidth). Right: in LT 3 (full bandwidth). dB, decibels; EQ, equalization; Hz, hertz.

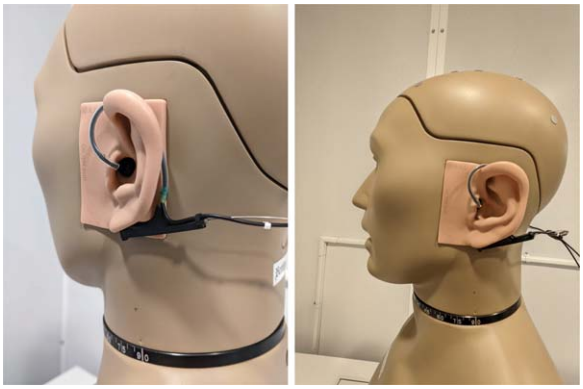

Fig. 2. Binaural microphone placed at the ear canal reference point (ECRP) of a KEMAR head and torso simulator.

the context of spatial audio rendering, the *flat* target is often defined near the entrance of the ear canal, which is a more accessible point when performing measurements on human subjects, such as for obtaining head-related transfer functions (HRTFs) or binaural room impulse responses (BRIRs) [16]. In this study, we performed the measurements with a pair of Brüel and Kjær (B&K) Type 4101-B binaural microphones, which were placed approximately 1 cm inside the ear canal, therefore ensuring that binaural cues were preserved (see Fig. 2). This measurement point is hereafter referred to as ear-canal reference point (ECRP). Thus, the *Harman* target had to be "translated" from DRP to ECRP, as described next.

The HpTF of the Audeze was first measured at the DRP of a KEMAR head and torso simulator (GRAS), using its internal microphones and ear canal simulators (GRAS RA0045), and then at the ERCP using the B&K microphones, which were the same ones that were later used for measuring HpTF on all participants in the listening tests. A HpEQ filter was generated to match the *Harman* target [15], as described in detail in SEC. 2.2. The measurements taken during this process are illustrated in the top row plots of Fig. 3. This procedure was repeated with the custom headset. The bottom row of Fig. 3 shows how, for the custom headset, the *Harman* target had to be limited in bandwidth due to the limitations of the built-in loudspeakers.

It is worth noting that the main difference between the ECRP-translated *Harman* target and the original one measured at the DRP is a missing peak around 3 kHz. This is explained by the fact that placing the microphone at the entrance of the semi-occluded ear canal limited the influence of the ear canal resonance on the measured frequency response [1].

1.5 Audio Material and Content Types

Four audio tracks (*material*) were used in the listening tests, as shown in Table 2. The three musical tracks were chosen as they had been proven to provide consistent preference ratings among subjects in previous listening tests [17]. All tracks were taken from original compact disks and re-sampled to a rate of 48 kHz.

In the listening tests, all audio material was presented as one of two content types: (i) *stereo*, generated by convolving the dry audio track with an HpEQ filter, or (ii) *spatial*, generated by convolving the dry audio track with a pair of individual BRIRs and with an HpEQ filter.

1.6 Listening Tests

Three listening tests were performed in order to investigate the research questions (a summary of the conditions tested in each listening test is given in Table 1):

1) Listening test 1 (LT 1): to assess whether the preferred target depended on the audio content type in the custom headset case, by testing five representative targets. A preliminary analysis of its results was presented in a previous paper [14].
2) Listening test 2 (LT 2): similar to LT 1 (also uses the custom headset) but a different set of targets is used with reduced variance between them to further ex-

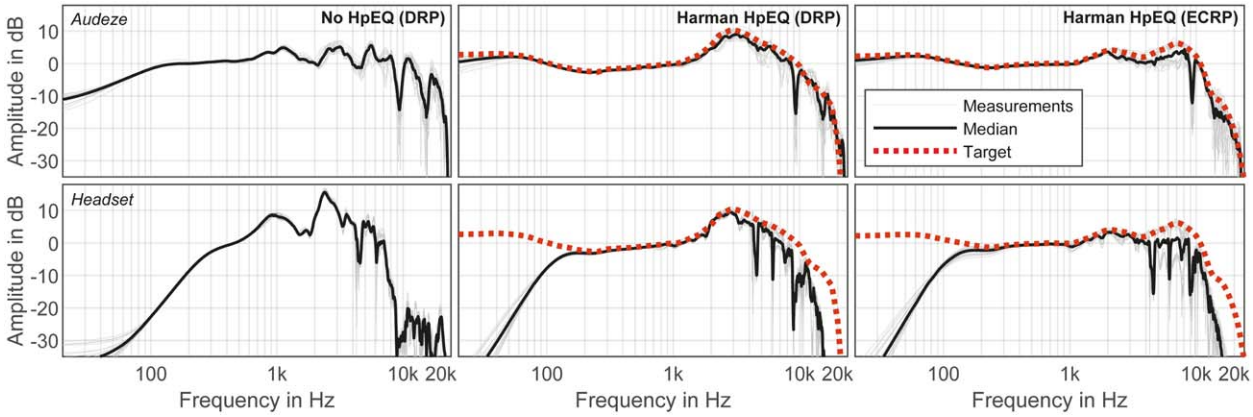

Fig. 3. Headphone transfer function (HpTF) magnitude response of the Audeze headphones (top row) and the custom headset (bottom row) with different headphone equalization (HpEQ) filters applied, measured on KEMAR at different reference points. Left: without HpEQ, measured at the eardrum reference point (DRP). Middle: after equalizing to the *Harman* target, measured at the DRP. Right: after equalizing to the *Harman* target, measured at the ear canal reference point (ECRP). Each plot shows 20 measurements (10 per ear), the median magnitude response, and the *Harman* target for the corresponding reference point.

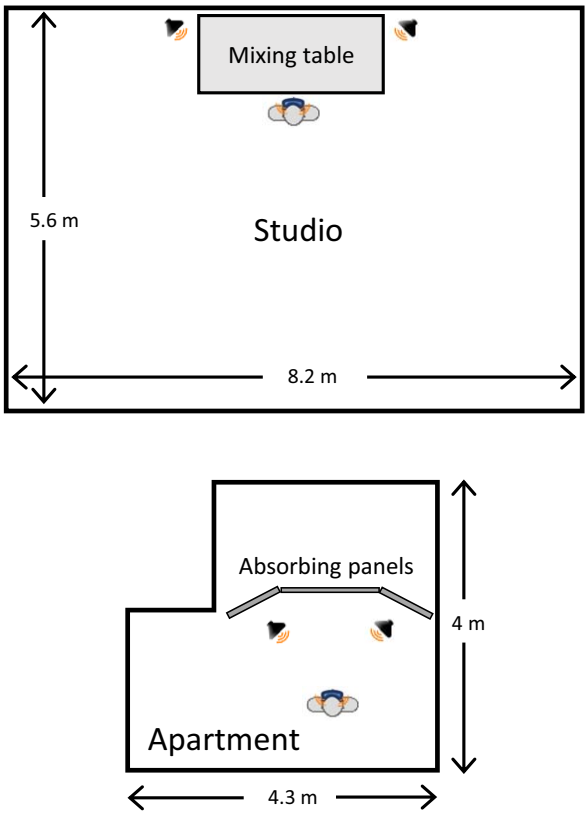

Fig. 4. Illustrations of the listening test rooms.

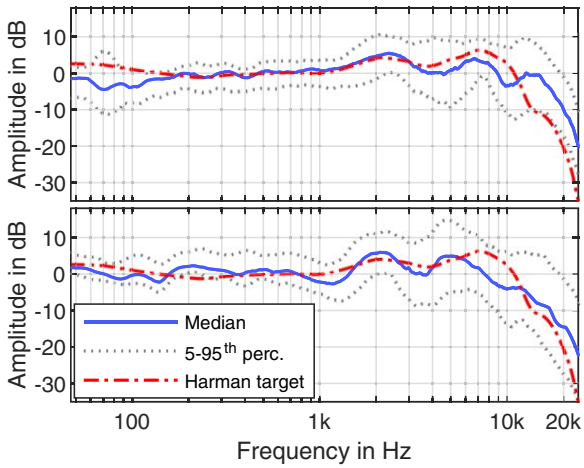

Fig. 5. Median, 5th, and 95th percentiles of binaural room impulse response (BRIR) magnitude spectrum across listeners, with a third-octave smoothing applied. Top: studio (24 measurements × 2 ears × 2 loudspeakers). Bottom: apartment (20 measurements × 2 ears × 2 loudspeakers). The *Harman* target, measured at the ear canal reference point (ECRP), is shown for comparison.

plore how target preference may depend on content type.

3) Listening test 3 (LT 3): similar to LT 2 (same set of targets) but using the Audeze headphones and two different reproduction bandwidths, which allowed us to evaluate the effect of the bandwidth and, indirectly, the effect of the headphones on the choice of preferred target and its dependence on content type.

2 METHODS

2.1 Listening Test Setup and Measurements

The listening tests were conducted in two different rooms, illustrated in Fig. 4. The first room (*apartment*) had an asymmetrical shape and was empty of furniture except for the test equipment and a set of portable absorbent panels. It had a reverberation time of $T30_{[400\,Hz\,-\,1250\,Hz]}$ = 499 ms. Two Genelec 8331A loudspeakers were placed in a stereo setup as defined by the ITU-R recommendation BS.2159 [18] and equalized flat at the listener's head location using an omnidirectional microphone (without the head present) following the procedure described by Olive et al. [11]. The second space (*studio*) was a control room in a recording studio ($T30_{[400\,Hz\,-\,1250\,Hz]}$ = 198 ms), hosting a pair of Focal SM-9 loudspeakers in a stereo setup, calibrated by a professional audio engineer.

Individual HpTFs, defined as the transfer functions between each headphone channel and the corresponding listener's ECRP, were generated for all the listeners from a single measurement performed at the beginning of the listening test, with a signal-to-noise ratio (SNR) of at least 60 dB, measured as the amplitude difference between the peak and the noise floor. After the measurements, the participants were asked not to touch or move the headphones in order to minimize potential variations in the HpTF. Generic HpTFs were measured on the KEMAR head as an upper variance limit of 10 measurements (SNR > 90 dB), according to Masiero and Fels [19], to avoid capturing high-frequency notches, which usually differ between listeners or when headphones are repositioned.

To provide convincing spatial audio content, individual BRIRs, defined as the transfer functions between each of the room's loudspeakers and each of the listener's ECRPs, were measured right after the HpTF measurements. BRIRs displayed an SNR of approximately 60 dB, measured as the amplitude difference between the peak and the noise floor. They were windowed at 500 ms (apartment) or 300 ms (studio) and subjected to a de-noising procedure as proposed by Cabrera et al. [20] to ensure a constant decay rate as the signal envelope approached the noise floor. An overview of the BRIR measurements is shown in Fig. 5.

The logarithmic sweep method [21] was used to obtain all HpTFs and BRIRs, using sweeps between 10 and 24000 Hz as excitation signals. The measurement hardware consisted of the aforementioned B&K microphones, a Brüel and Kjær 1407-A-002 signal conditioner and a RME Fireface UCX audio interface.

2.2 HpEQ

In order to obtain the desired targets from the headphones, minimum-phase HpEQ filters were calculated. This was done by frequency-domain division between the measured HpTF [H(ω)] and the target [T(ω)]. Frequency-dependent regularization was applied to prevent excessive amplification at frequencies for which the magnitude of the HpTF |H(ω)| was low (the reader is referred to the work by

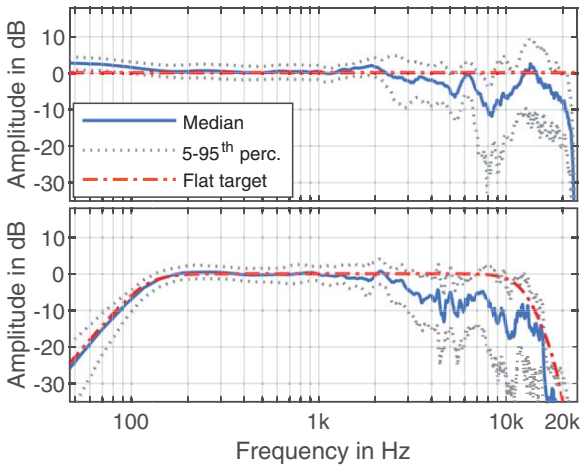

Fig. 6. Median, 5th, and 95th percentiles of headphone transfer function (HpTF) magnitude spectrum across listeners, after applying generic flat headphone equalization (HpEQ) (based on KEMAR measurements). Top: Audeze (30 measurements × 2 channels). Bottom: custom headset (44 measurements × 2 channels).

Kirkeby and Nelson [22] for more details on this technique and to Schärer and Lindau [23] for a perceptual evaluation of it). In this work, a regularization approach with automatic parameter adjustment was used, similar to one previously used by the present authors [7], which is based on the method proposed by Bolaños et al. [24]. The main novelty here is that T(ω) was taken into account when performing the regularized inversion, which guarantees that the resulting filter will meet the specified requirements, such as maximum gain and bandwidth. Thus, the regularized HpEQ filter EQ(ω) was defined as

$$\mathrm{EQ}(\omega) = \frac{\left[\frac{\mathrm{H}(\omega)}{\mathrm{T}(\omega)}\right]}{\left|\frac{\mathrm{H}(\omega)}{\mathrm{T}(\omega)}\right|^2 + \alpha + \sigma^2(\omega)}\mathrm{D}(\omega), \tag{2}$$

where D(ω) is a modeling delay to ensure causality, α defines the maximum amplification of the filter, and $\sigma^2(\omega)$ is a regularization factor defined as the negative deviation of the HpTF from a version where the magnitude of notches is reduced (i.e., smoothed along the frequency axis). In this work, it was defined that $\alpha = 2.5 \cdot 10^{-3}$ for a maximum amplification of 20 dB, and

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

where $\hat{\mathrm{H}}(\omega)$ is the result of applying fourth-octave-smoothing to H(ω). The reader is referred to [7, 24] for more information about the adjustment of the regularization parameters. Finally, all HpEQ filters (both generic and individual) where transformed to minimum-phase $[\mathrm{EQ}_{\mathrm{mp}}(\omega)]$, according to [25]:

$$\mathrm{EQ}_{\mathrm{mp}}(\omega) = |\mathrm{EQ}(\omega)| e^{-j \cdot \mathrm{Im}(\mathrm{Hilbert}(\ln(|\mathrm{EQ}(\omega)|)))}, \tag{4}$$

where Im(·) is the imaginary part, Hilbert(·) is the Hilbert transform, and ln(·) is the natural logarithm. This ensured a fair comparison between individual HpEQ filters and generic ones, which had their phase information removed as part of the averaging process as described in [19].

2.3 Listening Test Paradigm

In order to perceptually evaluate the different targets in terms of listener preference, a double-blind listening test was employed. The test was based on the multiple stimulus test with hidden reference and anchor (MUSHRA) defined in the ITU-R recommendation BS.1534 [26], and employed a similar interface, training procedure, and rating system. An important difference between the current paradigm and MUSHRA was that neither a hidden reference nor anchors were included here, to avoid introducing bias on listeners' ratings, given the highly subjective nature of the test. A similar paradigm has been used in previous studies on listener preference by Olive et al. [11–13].

Listeners were seated in front of a computer screen, holding a keyboard, and wearing the headphones being tested. In each trial, they were presented with five versions of the same audio material, each equalized to a different target, through the headphones. Listeners were then asked to rate each version according to their preference, using a graphical user interface (see Fig. 7), which contained one slider per target. The sliders were arranged in a different random order on every trial. The rating scale ranged between 0 and 100, with 5 point increments, and semantic labels were indicated every 20 points (from *Really Dislike* to *Really Like*) [13]. It was possible to seamlessly switch between stimuli as needed, and a time limit was not given to the listeners. The audio material had an approximate length of 10 s and looped automatically.

In pilot studies, listeners were asked to rate spatial, timbral, and overall quality in separate test sessions. However, it was found that results for the three metrics were often highly correlated, which may have been due to the three percepts being indeed correlated or perhaps to listeners being overwhelmed or fatigued by the length and complexity of the task and ending up rating according to their overall preference. In any case, it was decided to simplify the test by using a single global metric (preference) and performing informal post hoc interviews with the listeners to detect any unusual rating strategies.

Although the recommendation for MUSHRA tests is not to exceed 12 test signals (in this case, targets) per trial [26], it was found in pilot studies that using more than five signals already led to long listening tests and subsequent fatigue, due to the high number of test conditions. Therefore, it was decided to set the number of targets per trial to five in all listening tests. At the beginning of the test, listeners were required to complete a training stage in which they became familiar with the test equipment, grading scales, and all the sound excerpts under test, as recommended in [26].

Listeners could take part in two or all three listening tests, but they were required to do so on different days to avoid fatigue. The tests had a complete block design, meaning that listeners evaluated each combination of variables (i.e., *material/content* for LT 1 and 2, or *material/bandwidth* for LT 3) twice. Therefore, a full listening test consisted of

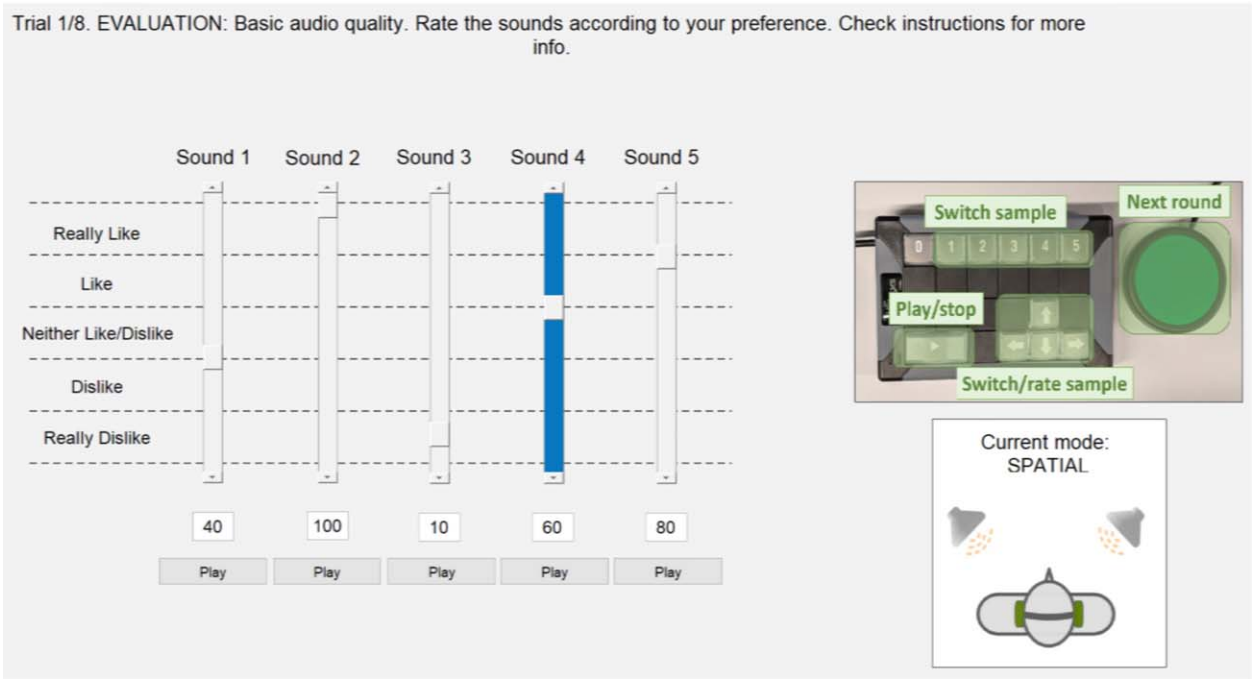

Fig. 7. Graphical user interface.

16 trials, which were divided in 2 blocks with a rest break between them.

A total of 38 listeners (25 male and 13 female) participated in the listening tests, of which seven were under 25 years old, 19 were in the 25–34 years range, six were in the 35–44 years range, and six were in the 45–54 years range. Listeners were recruited from two pools: Facebook employees (Redmond, WA, USA) and external naive listeners. As a pre-screening step, external listeners were selected for having achieved the best relative results in an audiometry test from a previous experiment. All listeners from the employee pool reported good hearing, except one person with mild loss above 8 kHz. Any listeners who performed poorly in the training stages (e.g., repeatedly failed to detect a difference between a reference signal and a low-passed version) would have been pre-screened as well, but this circumstance did not happen.

2.4 Analysis Approach

Our data were collected with a discrete ordered rating scale with 21 options (from 0 to 100 in 5 step increments). Because these data were bounded by the first and last options and were discrete rather than continuous, we modeled the responses ordinal data. Analysis of variance (ANOVA/simple linear regression) assumes data are unbounded, continuous, and spherical. As our data did not meet these assumptions, we analyzed the data with multilevel ordinal regression under a Bayesian framework to account for the discrete/ordinal nature of the responses (see [27] and [28]. Analyzing the data as ordinal avoids possible systematic errors, such as: false alarms (i.e., detecting an effect where none exists, Type I errors); failure to detect effects (i.e., loss of power, Type II errors); and inversions of effects, for which treating ordinal data as metric indicates the opposite ordering of means to the true ordering of means [27].

The model of LT 1 data included *target*, *material*, *content* and all interaction terms as population-level effects while *subject*, *trial*, *room*, and *subject group* were taken as varying (group-level) effects with correlation estimates for target, material, and content within subjects. The data from LT 2 and LT 3 was combined, both in order to estimate the effects of each headphone type and bandwidth and to pool variance estimates across the two experiments. The model of LT 2 and LT 3 included *target*, *material*, *headphone*, *bandwidth*, *content* and all interaction terms as population-level effects, and *subject*, *trial*, *room*, and *subject group* as varying (group-level) effects, with correlation estimates for target, material, headphone, bandwidth, and content within subjects. The multilevel nature of our model facilitated partial-pooling of group-level data and thus parameter estimates. With partial pooling, the probability of each response choice is modeled for each listener and the data for all participants also informs the estimates for participant [29].

3 LISTENING TEST 1

3.1 Description

The goal of the first listening test (LT 1) was to evaluate listener preference for several relevant targets in the custom headset case and to assess whether the ratings were affected by the audio content type (spatial or stereo). The following targets were evaluated (see also Table 1 and Fig. 1): (i) *ind. flat*, (ii) *gen. flat*, (iii) *Harman*, (iv) *no EQ*, and (v) *2 Harman*. As discussed in SEC. 1, the first three targets were chosen as they are the typical recommendations for the reproduction of spatial and stereo content, *no EQ* was added

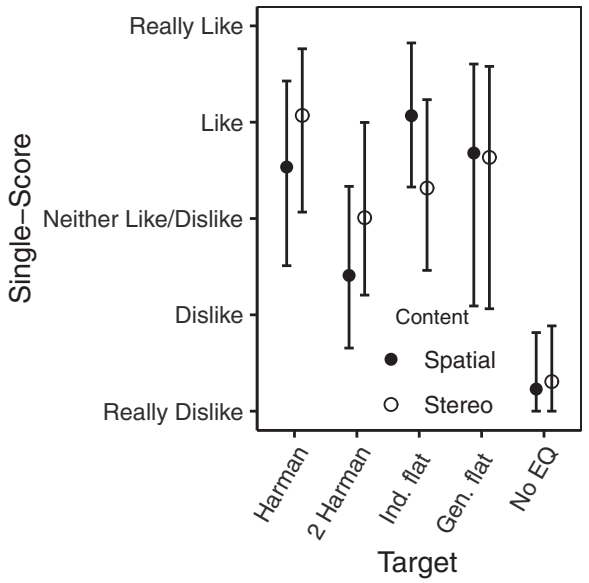

Fig. 8. Results from listening test 1, depicted as single-score ratings for each target and content type. Points represent the median single-score rating and error bars depict the 89% highest density credible interval. EQ, equalization; Gen., generic; Ind., individual.

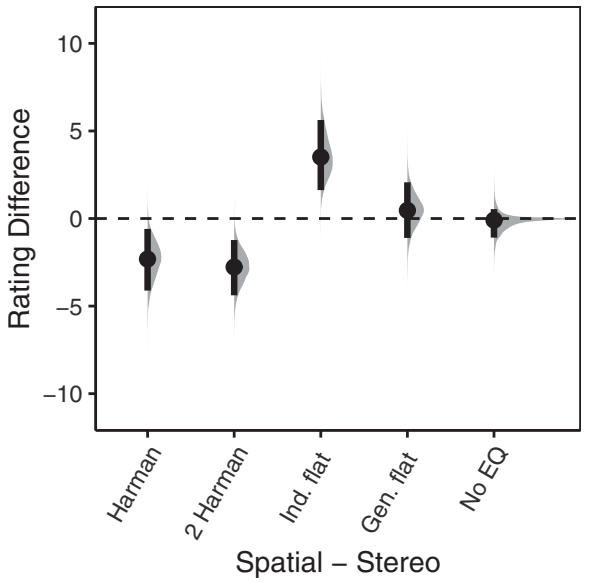

Fig. 9. Differences in rating between spatial and stereo content in listening test 1. Points represent median probability differences; bars represent the 89% highest density credible interval surrounding the means; and shaded regions indicate the posterior distribution of differences. Points and their respective intervals that do not include zero are statistically significant differences. Positive values indicate higher rating for spatial content and negative values indicate higher rating for stereo content. EQ, equalization; Gen., generic; Ind., individual.

as a representative HpTF of a VR headset without HpEQ, and *2 Harman* was included as a potential "perceptual intermediate" between *Harman* and *no EQ*—in the sense that it emphasized mid frequencies (between 1 and 5 kHz) more than the former but less than the latter. The custom headset was employed as headphones, and the limited-bandwidth targets were used. Both spatial and stereo content types were evaluated. Finally, each listener conducted the test in one of the two rooms (apartment or studio). LT 1 and a preliminary analysis of its results were presented in a previous paper [14].

3.2 Results

Of the pool of 38 listeners, 21 participated in the first listening test. The mean session time (not including breaks) across listeners was approximately 22 min, or 81 s per trial.

Median single-score ratings for each target and content together with 89% credible intervals are shown in Fig. 8. The 89% credible interval is computationally more stable relative to a 95% interval [30]. McElreath [31] suggested that 89% makes potentially more sense because 89 is "the highest prime number that does not exceed the already unstable 95% threshold." A quick inspection reveals that *no EQ* obtained the lowest ratings regardless of the content, followed by *2 Harman*, whereas the other three targets obtained comparatively higher ratings. The trends indicate that *ind. flat* was the highest rated target for spatial content, whereas *Harman* was the highest rated target for stereo content, which is in agreement with the initial hypotheses. However, the difference in rating between those two and *gen. flat* was ultimately not statistically significant due to the high variance of the data.

To follow up this analysis, Fig. 9 shows the difference between spatial and stereo ratings, separately for each target. It can be seen that *ind. flat* was rated significantly higher for *spatial* content than for *stereo* content, whereas *gen. flat* displayed a similar trend, but the difference was not significant. On the other hand, *Harman* and *2 Harman* were significantly preferred with *stereo* content. Finally, *no EQ* was not rated significantly higher for either content type. This supports the hypothesis that targets that approach *Harman* are preferred for stereo content, whereas those that approach *flat* are preferred for spatial content. The interactions with other variables such as material, subject group or room did not provide any meaningful findings and are not reported here for brevity.

3.3 Discussion

These results indicate that the *ind. flat* target was the preferred choice for spatial content, which is in line with previous studies that showed that an individually calibrated flat HpTF is the optimal choice for binaural audio reproduction [4, 5]. On the other hand, *Harman* seemed to obtain higher ratings than the other targets for stereo content, supporting the findings of Olive et al. [11–13]. Furthermore, a significant effect of the content type was observed, with *ind. flat* obtaining higher ratings for spatial content than for stereo content and vice versa for *Harman*. These results support the hypothesis that audio content type has an effect on the target preference, at least when individual HpEQ is employed. However, this effect was less evident when generic HpEQ was employed: although *gen. flat* displayed similar trends to *ind. flat*, the rating differences between content types were not significant.

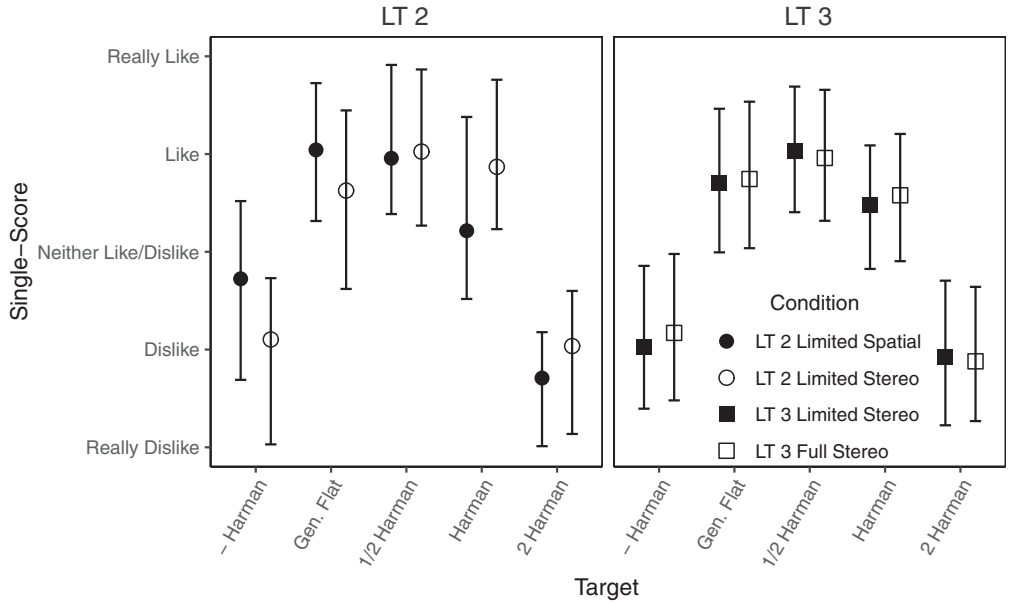

Fig. 10. Results from listening test (LT) 2 and LT 3, depicted as single-score ratings for each target and content type (LT 2) or bandwidth (LT 3). Points represent the median single-score rating and error bars depict the 89% highest density credible interval. Gen., generic.

A possible explanation for this apparent lack of effect in the generic case is that the presence of the optimal target in the spatial audio condition (*ind. flat*), as well as the anchor target (*no EQ*), both of which were likely to produce extreme ratings, led listeners to rate other targets as being similar. Another explanation is that *gen. flat* and *Harman* are similar enough to each other from a perceptual point of view, enough so to obtain similar preference ratings. On the other hand, *2 Harman*, which shows a larger spectral deviation from the *gen. flat* than *Harman*, did produce a significantly lower rating. Therefore, it is possible that listeners tolerate a certain amount of spectral deviation from the optimal target, and preference ratings drop significantly when this tolerance is surpassed. This was explored in the subsequent listening tests.

4 LISTENING TESTS 2 AND 3

4.1 Description

From the results from LT 1, it was hypothesized that, when generic HpEQ is employed, listeners may tolerate a certain amount of spectral deviation from the optimal target, and preference ratings drop only after that threshold is surpassed.

The second listening test (LT 2) was identical to LT 1 except that a different set of targets was used. The low-quality anchor (*no EQ*) was eliminated and so was *ind. flat*. In general, the targets in LT 2 displayed smaller variance between them that the ones in LT 1, which might help in reducing the variance in the ratings and lead to more insights into the research question. The following targets were tested in LT 2: (i) *−1 Harman*, (ii) *(gen.) flat*, (iii) *1/2 Harman*, (iv) *Harman*, and (v) *2 Harman*. As in LT 1, the custom headset was employed as headphones, the limited-bandwidth targets were used, and both spatial and stereo content types were evaluated.

The third listening test (LT 3) explored the effect of the headphone's reproduction bandwidth on the preference ratings. The goal was to gain insight on whether the limitations of open-ear hardware had a significant impact on listener preference and its dependence on audio content type. In LT 3, the Audeze were employed as headphones and the same set of targets from LT 2 were evaluated, except that both full and limited bandwidths were assessed. Only stereo content was employed in LT 3.

4.2 Results

For LT 2, 23 listeners voluntarily participated, and the mean session time (not including breaks) was approximately 19 min, or 72 s per trial. For LT 3, 28 listeners voluntarily participated, and the mean session time (not including breaks) was also 19 min, approximately.

Results of both listening tests are displayed in Fig. 10, which shows the single-score ratings for each target and content (LT 2) or bandwidth (LT 3). It can be clearly seen that, like in LT 1, listener preference strongly depended on the target. In this case, *2 Harman* and *−1 Harman* obtained the lowest ratings overall. No significant differences were observed between the single-score ratings of *Harman*, *flat*, and *1/2 Harman*. Data from LT 2 suggest that there might be an effect of audio content type on the choice of target, given that the preference ratings show slightly different trends for the spatial and stereo cases. However, it is hard to determine if there exist significant differences from Fig. 10 alone, so further analysis is required. On the other hand, LT 3 data suggest that the reproduction bandwidth had little effect on the choice of target, with both full and limited bandwidths displaying almost identical trends. The inferential analysis

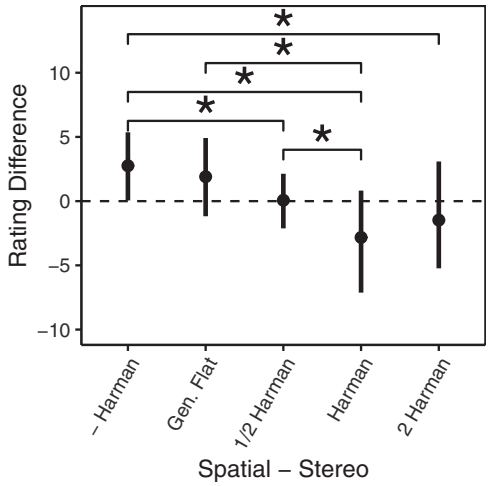

Fig. 11. Differences in rating between spatial and stereo content in listening test 2 (LT 2). Points represent median probability differences; bars represent the 89% highest density credible interval surrounding the means. Positive values indicate higher rating for spatial content and negative values indicate higher rating for stereo content. Brackets and asterisks indicate statistically significant differences. Gen., generic.

confirmed that none of the target's ratings were significantly affected by chosen reproduction bandwidth (full vs. limited).

The difference in ratings between spatial and stereo content for each target of LT 2 is displayed in Fig. 11. A trend can be observed in which *flat* and the targets that are closer to it (*−1 Harman*) are preferred for spatial content, whereas, in contrast, *Harman* and *2 Harman* are rated slightly higher for stereo content. Most notably, we observe a statistically significant difference between *flat* and *Harman* (other significant differences are indicated in Fig. 11). Finally, *1/2 Harman* falls approximately in the middle, not being clearly preferred for either audio content type. These results align well with the initial hypotheses, although none of the targets except − *Harman* showed a rating difference significantly different from zero, suggesting that the audio content type may not have a significant effect on the rating.

Nevertheless, significant effects were found when analyzing the data per material, as shown in Fig. 12. In particular, *−1 Harman* and *flat* are rated significantly higher for spatial content regardless of the material; *Harman* and *2 Harman* are rated significantly higher for stereo content for all material except speech (SP); and *1/2 Harman* is not significantly preferred for either content type. It is hypothesized that these significant effects did not arise in the initial analysis because the data were collapsed across different independent variables, which increased the uncertainty in the estimates due to potential interactions between said variables. For instance, once the data were split per material, it was observed that the SP produced less significant effects than other materials (JW, SD, GA; please refer to Table 2 for more information), as observed in Fig. 12. Informal post hoc interviews revealed that some listeners might have given higher ratings to *Harman* and *2 Harman* for the speech material regardless of the content type, simply because these targets emphasized speech frequencies, slightly improving intelligibility even though the result was less authentic.

4.3 Discussion

Results from LT 2 showed that, once the variable of generic vs. individual HpEQ was removed from LT 1, a significant interaction between target preference and the audio content type could be observed. Although the single-score ratings did not reveal any significant differences in overall preference between *Harman*, *1/2 Harman* and *flat*, probably due to interlistener variance, significant differences were observed between the preference ratings for spatial and stereo content within each target when analyzed separately per material. From observing these results, it can be said that listeners' preference of target depends on the audio content type, even when generic HpEQ is employed. However, this effect of content type might be less than that observed in LT 1, in which individual HpEQ was considered. This difference might be explained by the fact that employing generic HpEQ may introduce errors of comparable magnitude to the differences between the evaluated targets, leading to higher variance in ratings among listeners.

This explanation is supported by Fig. 6, which shows individual HpTF measurements after applying the *gen. flat* HpEQ filter. It can be seen that interlistener variations are relatively large for frequencies above 2 kHz, which is in line with results published in previous literature [6, 16] and sometimes even larger than the differences between *flat* and *Harman* targets themselves (cf. Fig. 1). In other words, some listeners may perceive an HpTF that differs considerably with the intended target and potentially have a worse experience than other listeners—possibly, those with an anatomy more similar to the KEMAR head, where the generic measurements were taken. Furthermore, it is evident that the median measurement after applying KEMAR-based equalization actually deviates considerably from the intended target (flat). This observation complies with previous research by Lindau and Brinkmann [6] and is an indication that KEMAR may not be a good representative of this population.

Moreover, the generic HpTF was calculated as an upper variance limit of several measurements, as described in SEC. 2, which may have introduced a positive bias on its high-frequency magnitude response and, therefore, added to the overall bias of the KEMAR-based HpEQ filters. If, instead, a generic target based on the actual measured listeners (e.g., the mean magnitude response of the individual HpTFs) were used, the experience of the average listener should improve. However, even in this case, the issue of variability across listeners would still exist, as it is an inherent limitation of generic HpEQ. Future work could investigate the relationship between a listener's rating preference and the amount of deviation of their individual HpTF from the generic one used to generate the HpEQ filters.

Another factor that was initially taken into account was the limited reproduction bandwidth imposed by using the

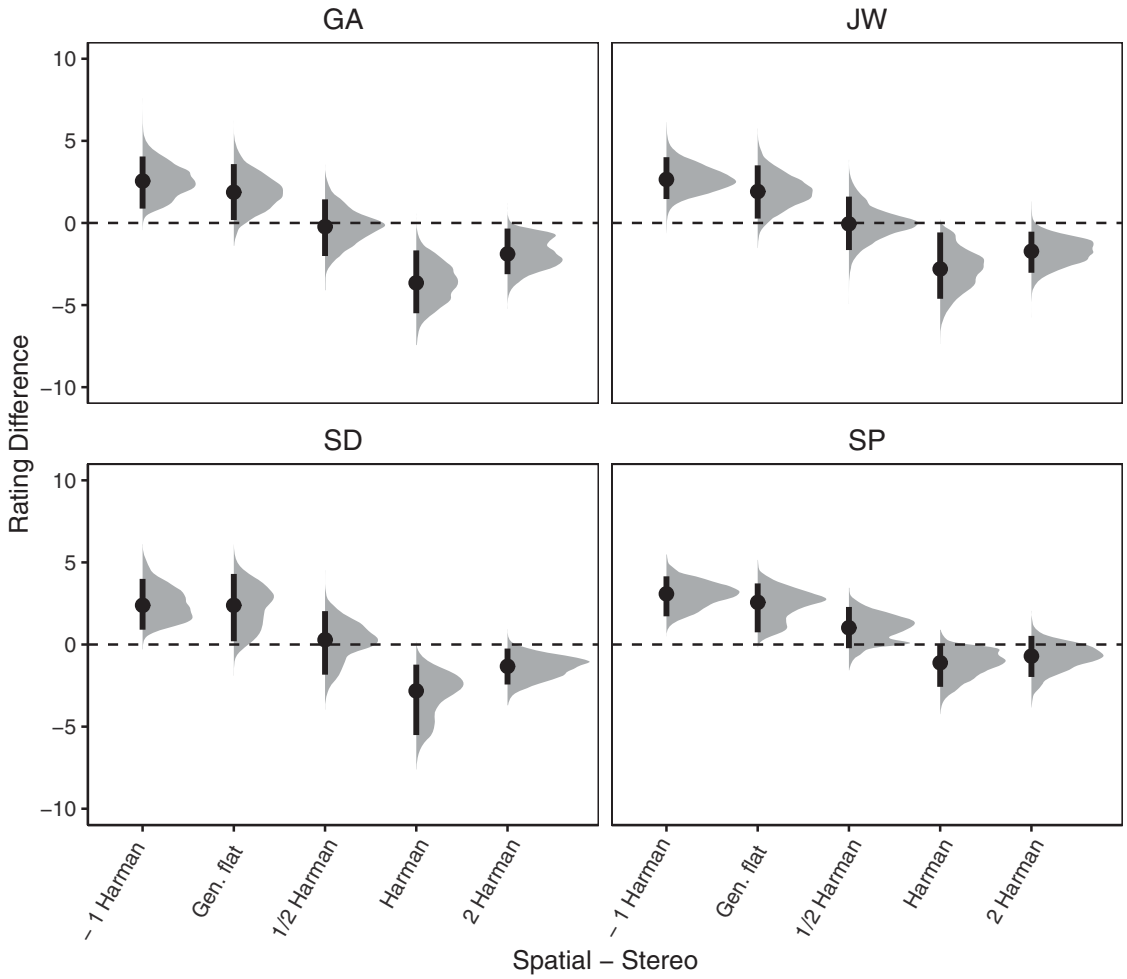

Fig. 12. Same as Fig. 11, but data are separated per audio material. Intervals that do not include zero indicate that the rating differences are statistically significant. Titles signify audio tracks according to the keys in Table 2.

custom VR headset as a playback device. This limitation meant, for instance, that the presented *Harman* target was missing some characteristic features from the original one (defined between 20 and 20,000 Hz), such as a low-frequency boost (<100 Hz) and a roll-off above 10 kHz. However, the results of LT 3 indicate that the reproduction bandwidth did not affect the preference of target for the tested conditions, as the ratings for full-bandwidth targets were similar to the ones where the frequency range was limited. This suggests that the results from LT 1 and LT 2 were not affected by the limited bandwidth of the custom headset and therefore could be applicable to systems without the said constraints.

Note that if the full-bandwidth content had been compared with the limited-bandwidth directly (i.e., in the same trial), we would most likely observe a strong preference bias towards the former. However, it is worth recalling that the research question was not whether listeners prefer full-bandwidth content but whether the relative HpTF preference is significantly affected by the reproduction bandwidth, which the results suggest it is not. Similarly, the rating trends for the Audeze headphones and the custom headset were found to be almost identical for stereo content, which suggests that the results reported in this study are independent of the headphones themselves. However, in order to generalize the results to any binaural reproduction system, other factors such as crosstalk or distortion should be explored separately.

5 SUMMARY

This study addressed the issue of the effect of audio content type on listener preference for the target HpTF. Based on previous studies, it was hypothesized that a flat target would be preferred for spatial content, whereas one which mimics the magnitude response of a loudspeaker system in a listening room (e.g., *Harman* target) would perform better for non-spatial stereo content. An important aspect of the evaluation was prioritizing test conditions which were relevant for the custom headset case (as a representative of a typical AR/VR product). For this reason, a custom prototype VR headset was used as one of the binaural playback systems. The outcomes of this study can be summarized as follows:

1) When individual headphone equalization was used, a clear effect of audio content type was observed in

listeners' preference for headphone response. The flat target was generally preferred for spatial (binaural) content and the *Harman* target, for stereo content.

2) This effect was also observed for generic headphone equalization. However, it was smaller than in the individual equalization case, due to interlistener variations in HpTF.
3) The reproduction bandwidth of the system did not have an effect on the preference of HpTF, which allows for the generalization (to some extent) of these results to systems without bandwidth limitations typical of open-ear headsets.

6 REFERENCES

[1] H. Møller, "Fundamentals of Binaural Technology," *Appl. Acoust.*, vol. 36, nos. 3–4, pp. 171–218 (1992). https://doi.org/10.1016/0003-682X(92)90046-U.

[2] S. E. Olive, "A New Reference Listening Room for Consumer, Professional and Automotive Audio Research," presented at the *126th Audio Engineering Society Convention* (2009 May), paper 7677.

[3] IEEE, "IEEE Standard for Translating Head and Torso Simulator Measurements from Eardrum to Other Acoustic Reference Points," *Standard 1652-2016* (2016 Nov.). https://doi.org/10.1109/IEEESTD.2016.7755724.

[4] D. Pralong and S. Carlile, "The Role of Individualized Headphone Calibration for the Generation of High Fidelity Virtual Auditory Space," *J. Acoust. Soc. Am.*, vol. 100, no. 6, pp. 3785–3793 (1996 Dec.). https://doi.org/10.1121/1.417337.

[5] H. Møller, M. F. Sørensen, C. B. Jensen, D. Hammershøi and, "Binaural Technique: Do We Need Individual Recordings?" *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469 (1996 Jun.).

[6] A. Lindau and F. Brinkmann, "Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings," *J. Audio Eng. Soc.*, vol. 60, nos. 1–2, pp. 54–62 (2012 Jan.).

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

[8] F. E. Toole, "The Acoustics and Psychoacoustics of Headphones," in *Proceedings of the Audio Engineering Society 2nd International Conference: The Art and Technology of Recording* (1984 May), paper C1006.

[9] H. Møller, C. B. Jensen, D. Hammershøi, M. F. Sørensen and, "Design Criteria for Headphones," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 218–232 (1995 Apr.).

[10] G. Lorho, "Subjective Evaluation of Headphone Target Frequency Responses," presented at the *126th Audio Engineering Society Convention* (2009 May), paper 7770.

[11] S. E. Olive, T. Welti, and E. McMullin, "Listener Preferences for Different Headphone Target Response Curves," presented at the *134th Audio Engineering Society Convention* (2013 May), paper 8867.

[12] S. E. Olive, T. Welti, and E. McMullin, "The Influence of Listeners' Experience, Age, and Culture on Headphone Sound Quality Preferences," presented at the *137th Audio Engineering Society Convention* (2014 Oct.), paper 9177.

[13] S. E. Olive, T. Welti, and O. Khonsaripour, "A Statistical Model That Predicts Listeners' Preference Ratings of In-Ear Headphones: Part 2—Development and Validation of the Model," presented at the *143rd Audio Engineering Society Convention* (2017 Oct.), paper 9878.

[14] I. Engel, D. L. Alon, K. Scheumann, and R. Mehra, "Listener-Preferred Headphone Frequency Response for Stereo and Spatial Audio Content," in *Proceedings of the Audio Engineering Society International Conference on Audio for Virtual and Augmented Reality* (2020 Aug.), paper 1-5.

[15] S. E. Olive and T. Welti, "Factors That Influence Listeners' Preferred Bass and Treble Levels in Headphones," presented at the *139th Audio Engineering Society International Convention* (2015 Oct.), paper 9382.

[16] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Transfer Characteristics of Headphones Measured on Human Ears," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 203–217 (1995 Apr.).

[17] S. E. Olive, T. Welti, and O. Khonsaripour, "The Influence of Program Material on Sound Quality Ratings of In-Ear Headphones," presented at the *142nd Audio Engineering Society Convention* (2017 May), paper 9778.

[18] ITU-R, "Multichannel Sound Technology in Home and Broadcasting Applications," Rep. BS.2159-8 (2019 Jul.). https://itu.int/pub/R-REP-BS.2159-8-2019.

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

[20] D. Cabrera, D. Lee, M. Yadav, and W. L. Martens, "Decay Envelope Manipulation of Room Impulse Responses: Techniques for Auralization and Sonification," in *Proceedings of Acoustics 2011* (Gold Coast, Australia) (2011 Nov.).

[21] A. Farina, "Advancements in Impulse Response Measurements by Sine Sweeps," presented at the *122nd Audio Engineering Society Convention* (2007 May), paper 7121.

[22] O. Kirkeby and P. A. Nelson, "Digital Filter Design for Inversion Problems in Sound Reproduction," *J. Audio Eng. Soc.*, vol. 47, nos. 7–8, pp. 583–595 (1999 Jul.).

[23] Z. Schärer and A. Lindau, "Evaluation of Equalization Methods for Binaural Signals," presented at the *126th Audio Engineering Society Convention* (2009 May), paper 7721.

[24] J. G. Bolaños, A. Mäkivirta, and V. Pulkki, "Automatic Regularization Parameter for Headphone Transfer Function Inversion," *J. Audio Eng. Soc.*, vol. 64, no. 10, pp. 752–761 (2016 Oct.). https://doi.org/10.17743/jaes.2016.0030.

[25] A. V. Oppenheim, J. R. Buck, and R. W. Schafer, *Discrete-Time Signal Processing, Vol. 2* (Prentice Hall, Upper Saddle River, NJ, 2001).

[26] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems," Rep. BS.1534 (2015 Oct.). https://itu.int/rec/R-REC-BS.1534.

[27] T. M. Liddell and J. K. Kruschke, "Analyzing Ordinal Data with Metric Models: What Could Possibly Go Wrong?" *J. Exp. Soc. Psychol.*, vol. 79, pp. 328–348 (2018 Nov.). https://doi.org/10.1016/j.jesp.2018.08.009.

[28] P.-C. Bürkner and M. Vuorre, "Ordinal Regression Models in Psychology: A Tutorial," *Adv. Methods Pract. Psychol. Sci.*, vol. 2, no. 1, pp. 77–101 (2019 Feb.). https://doi.org/10.1177/2515245918823199.

[29] A. Gelman and J. Hill, *Data Analysis using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, Cambridge, UK, 2006).

[30] J. K. Kruschke, "Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan," (Cambridge, MA, Academic Press, 2014), 2nd ed.

[31] R. McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (London, UK, Chapman and Hall/CRC, 2020), 2nd ed.

THE AUTHORS


Isaac Engel


David Lou Alon


Kevin Scheumann


Jeff Crukley


Ravish Mehra

Isaac Engel received a B.Sc. in Electronic Systems Engineering and a Master's Degree on Telematics and Telecommunication Networks from the University of Málaga in 2015 and 2016, respectively. He is currently a doctoral candidate in the Dyson School of Design Engineering at Imperial College London, in the United Kingdom, where he researches spatial audio perception for audio Augmented Reality. In 2018 and 2019, he worked as a research intern at Facebook Reality Labs, where his research focused on headphone equalization.

•

David Lou Alon is a Research Scientist at Facebook Reality Labs (FRL) Research, leading the research on spatial audio technologies. He received his Ph.D. in electrical engineering from Ben Gurion University (Israel, 2017) in the field of spherical microphone array processing. His research areas include head-related transfer functions, spatial audio capture, binaural reproduction, and headphone equalization for VR and AR application.

•

Kevin Scheumann is a Technical Program Manager at Facebook Reality Labs (FRL) Research. He received a BA in Telecommunications with a focus on Media Design and Production from Indiana University Bloomington. His previous work spanned game design, analysis, and production, as well as hardware validation. His current work focuses on data collection, machine learning, and computer vision for audio research.

•

Through a decade as an industry scientist and university instructor in the field of communication sciences and disorders, Jeff Crukley grew fascinated with how we handled data and presented analyses. He now practices primarily as an applied statistician in collaboration with scientists/researchers. He has worked with academics in education, cognitive psychology, audiology, palliative care, speech-language pathology, surgery, and epidemiology. He has contributed to instrument development and validation, and his research contributions have led to changes in clinical teaching practices. He also continues to contribute to the med-tech industry, in which his work is integral to new and improved products and FDA approvals. He also mentors graduate students through his university appointments at University of Toronto and McMaster University. He completed a post-doctoral fellowship in Medical Biophysics and Neuroscience, a PhD in Hearing Science, an MSc in Communication Sciences & Disorders from Western University, and a BSc in Biology and Psychology from McMaster University. He is constantly enrolled in formal and informal continuing education and professional development in statistics and data science.

•

Ravish is the Director for Audio Research at Facebook Reality Labs (FRL) Research responsible for developing novel audio techniques to push the state-of-the-art for audio in VR and AR. He completed his Ph.D. in Computer Science at the University of North Carolina at Chapel Hill in the field of acoustics and spatial audio. In his doctoral work, he worked on novel physically based simulation techniques for simulating complex acoustic phenomena arising out of propagation of sound waves in large environments. His research interests span the fields of audio, acoustics, signal processing, and virtual and augmented reality. Dr. Mehra's work in acoustics and spatial audio has generated considerable interest in the audio community, and his sound propagation and spatial sound system has been integrated into virtual reality systems (Oculus HMD), with demonstrated benefits.