Audio Engineering Society

Conference Paper

Presented at the 2020 Conference on
Audio for Virtual and Augmented Reality
2020 August 17 – 19, Redmond, WA

Listener-Preferred Headphone Frequency Response for Stereo and Spatial Audio Content

Isaac Engel, David Lou Alon, Kevin Scheumann, and Ravish Mehra

Facebook Reality Labs, Redmond WA, United States

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

ABSTRACT

When spatial audio content is presented over headphones, the audio signal is typically filtered with binaural room impulse responses (BRIRs). An accurate virtual auditory space presentation can be achieved by flattening the headphones' frequency response. However, when presenting stereo music over headphones, previous studies have shown that listeners prefer headphones with a frequency response that simulates loudspeakers in a listening room. It is as yet unclear if headphones that are calibrated in such a way will be preferred by listeners in the context of spatial audio content as well. This study investigates how listeners' preferences for headphone frequency response may differ between stereo audio content and spatial audio content, which was rendered by convolving the same stereo content with in-situ-measured BRIRs of loudspeakers in a room.

1 Introduction

When presenting spatial audio content over headphones, it has been shown that optimal results are achieved when the headphone transfer function (HpTF) is individually equalized towards a flat response, as it promotes accurate reconstruction of idiosyncratic binaural and monoaural cues [1, 2, 3]. Recent studies have also shown that even when generic headphone equalization (HpEQ) is used, improvements in overall quality, coloration and externalization are gained [4, 5].

However, when commercial stereo audio content is reproduced over headphones, different target HpTFs may be preferable to a flat response. Møller et al. [6] suggested that, in such a case, the HpTF should match the frequency response of a loudspeaker system as received by a listener in free or diffuse fields. On this premise, the *Harman target* curve, based on acoustical measurements in a calibrated listening room, was proposed by Olive et al. [7]. The Harman target curve was evaluated by listeners who were presented with commercial stereo music through headphones, and was found to be preferable to other target curves, such as the free and diffuse field targets suggested by Møller et al. [6], and a variation of the latter proposed by Lorho [8]. Further studies by Olive et al. [9, 10] showed that the Harman target was preferred to the HpTFs of off-the-shelf headphones for commercial stereo music, by over 300 listeners.

From the reviewed literature, it seems that the choice of

target HpTF should depend on the type of audio content (spatial, stereo). However, to the best of the authors' knowledge, the effect of audio content on headphone target preference has not yet been studied explicitly. This question becomes increasingly relevant as more commercial devices, such as head-tracked headphones with spatial audio technology, virtual reality (VR) headsets and augmented reality (AR) glasses, are intended to play both spatial and non-spatial audio content. In this study, listeners' preferences for target HpTFs is evaluated through objective analysis and listening tests for stereo and spatial audio contents, focusing on the case of VR headset usage.

2 Methods

A listening test was designed to evaluate listeners' preferred target HpTF for spatial and stereo audio contents under various conditions, described in the following subsections.

The focus of this study is the case of the VR headset. A custom VR headset prototype with built-in loudspeakers was therefore chosen as "headphones" in the evaluation. This device provided realistic audio bandwidth limitations related to the loudspeakers' size, and to their off-ear location (which is common in commercial VR headsets). The headset had its frontal part removed so the listener could see through it, to reduce any visual biases that could be associated with a displayed virtual scene.

2.1 Evaluated target HpTFs

Five target HpTFs were selected for the evaluation: two that were previously recommended for spatial audio content, namely **individual flat** and **generic flat**; the **Harman** target [11], which was previously recommended for stereo content; the unequalized HpTF of the custom VR headset prototype (**no EQ**), which may be regarded as representative of a VR headset; and an "exaggerated" version of the Harman target (denoted as **2 x Harman**), which was preferentially rated, in an informal listening test, as intermediate between the **Harman** and **no EQ** HpTFs.

2.2 Audio material and content types

Four different audio materials were used to evaluate listeners' preference, including three music tracks and one speech sample, as listed in Table 1. The three music tracks were selected to be those that produced the most reliable ratings in a previous study by Olive et al. [12]. These could be presented as **stereo** content (dry audio convolved with a HpEQ filter), or **spatial** content (generated by convolving the **stereo** content with individualized BRIRs).

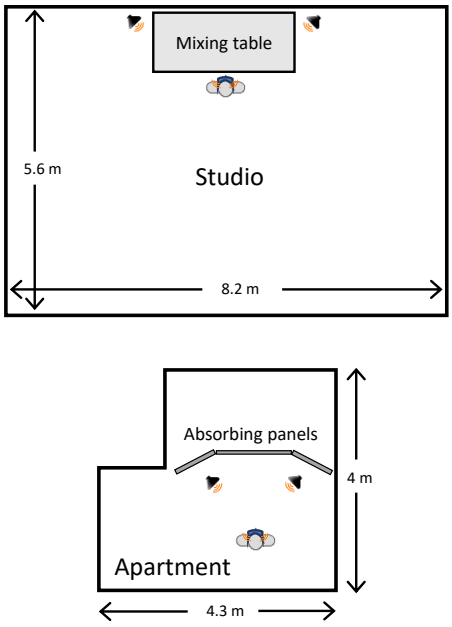

Fig. 1: Illustration of the listening test rooms.

2.3 Setup and hardware

Two rooms were used for the listening test, as shown in Fig. 1. The first one (**Apartment**) was an approximately 4 x 4.3 x 2.7 m room, empty of furniture except for several portable absorbent panels used as acoustical treatment, and had a reverberation time of $T30_{[400Hz-1250Hz]}$ = 499 ms. The second one (**Studio**) was a 8.2 x 5.6 x 3.1 m control room of a recording studio, which was treated to reduce reflections from the window and mixing table shown in Fig. 1, and had a reverberation time of $T30_{[400Hz-1250Hz]}$ = 198 ms. In both rooms, a stereo loudspeaker setup was placed in front of the listener. In the case of the Apartment, a pair of Genelec 8331A monitors was used, calibrated according to Olive et al. [7]. In the Studio, a pair of Focal SM-9 was used, and was manually tuned by a professional audio engineer.

Table 1: Material used in the listening test.

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

2.4 Measurements and equalization

In order to be able to apply individual HpEQ (for the **individual flat** target) and to conduct the objective analysis, individual HpTFs were measured for all participants at the beginning of the experiment. In addition, a generic HpTF was measured on a GRAS 45BC KEMAR head and torso simulator with KB5000/KB5001 ears. Measurements were performed via 2-second-long logarithmic sweeps between 10 and 24000 Hz [13]. In the case of the generic HpTF, an average of 10 measurements was generated according to Masiero and Fels [14]. A pair of binaural microphones, Brüel and Kjær (B&K) 4101-B, placed at the entrance of the ear canal and partially occluding it, was used in both the individual and the generic measurements.

To produce the desired target HpTFs, HpEQ filters were calculated by frequency-domain division between the target and the measured HpTFs. Regularization was applied to prevent excessive amplification outside the headphone frequency range, as well as inversion of narrow notches at high frequencies [5, 15]. All HpEQ filters were generated as minimum-phase.

In order to be able to generate realistic **spatial** audio content, individual BRIRs were measured for all participants from each of the two loudspeakers to both ears, using 8-second-long logarithmic sweeps between 10 and 24000 Hz. This was done immediately after measuring the HpTF, and using the same microphone positions. The BRIRs were denoised according to [16] to ensure a constant decay rate when approaching the noise floor. Participants were instructed not to touch the headset during the experiment to avoid modifying the HpTF due to repositioning.

2.5 Test procedure

The listening test had a double blind paradigm, similar to those used in previous related studies [7, 9, 10]. It was inspired by the MUSHRA (MUltiple Stimulus test with Hidden Reference and Anchor) paradigm [17], but did not employ a hidden reference or anchors, to avoid any potential bias in the judgment of the listeners.

In each trial, listeners were asked to evaluate five sound excerpts which were generated by applying the evaluated target HpTFs (in a random order) to the same audio material. Listeners were instructed to rank the sound excerpts based on their preference, using a rating scale from 0 to 100 (semantically labeled from *Really Dislike* to *Really Like*), as shown in Fig. 2. Each sound excerpt lasted 10 seconds approximately and looped automatically. Listeners used a custom keyboard to assign the ratings, stop and restart the audio playback, and seamlessly switch between the excerpts.

Before the test, listeners were required to complete two training sessions where they were introduced to the test equipment, the grading scales and the sound excerpts that they would evaluate, as recommended in [17].

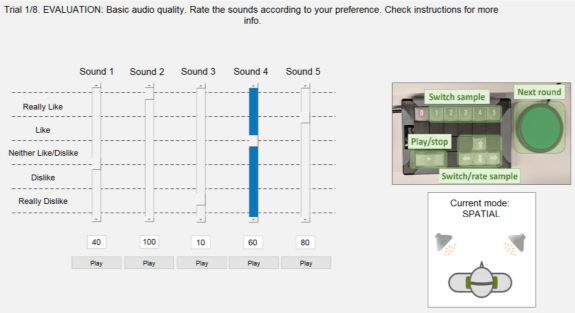

Fig. 2: Graphical user interface of the listening test.

3 Objective analysis

Different ways to define and measure target HpTFs have previously been suggested, such as the ear-Drum Reference Point (DRP) or the Ear canal entrance Reference Point (ERP) [6, 18]. To facilitate comparison

between the evaluated target HpTFs, all target HpTFs in this study are measured with respect to the ERP (semi-blocked with B&K binaural microphones as described in the previous section).

The **Harman** target was originally defined for the DRP of a head and torso simulator [7], and was therefore translated to the ERP. This was achieved by equalizing a pair of Audeze LCD-2 headphones (same model which was originally used to evaluate the **Harman** target [7]) to the **Harman** target at the DRP of a KEMAR head, and then measuring them again at the ERP of the same head with the B&K binaural microphones.

Figure 3 shows the HpTF of the custom VR headset after applying HpEQ filters for the **Harman** target at the DRP and at the ERP, as described in the previous section. Due to the bandwidth limitations which are introduced by the VR headset open-ear design, some of the features of the original **Harman** target are missing, such as a boost below 150 Hz and a roll-off at high frequencies, which fall outside the reproduction bandwidth of the headset.

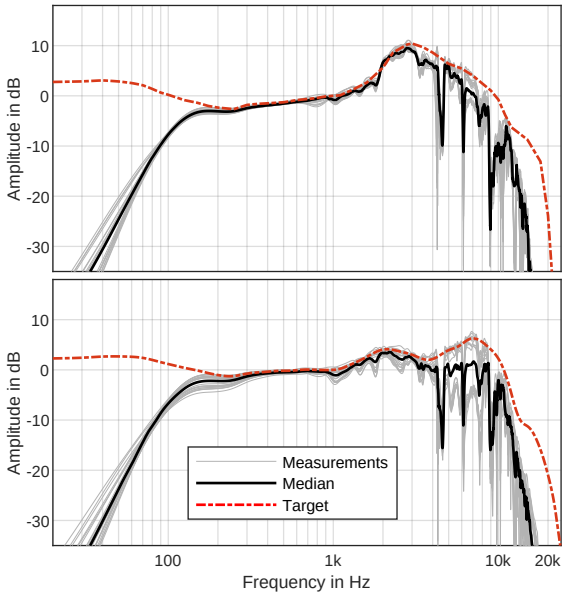

Fig. 3: HpTF of the custom VR headset after equalizing to the **Harman** target, measured at the DRP (top) and at the ERP (bottom). Each plot shows 10 KEMAR measurements on left and right ears, the median magnitude response and the target HpTF.

Figure 4 shows the magnitude response of the five evaluated target HpTFs. A 3rd order band-pass filter between 120 and 11300 Hz was applied to all target HpTFs to approximately match the reproduction bandwidth of the custom VR headset prototype that was used as headphones in the experiment.

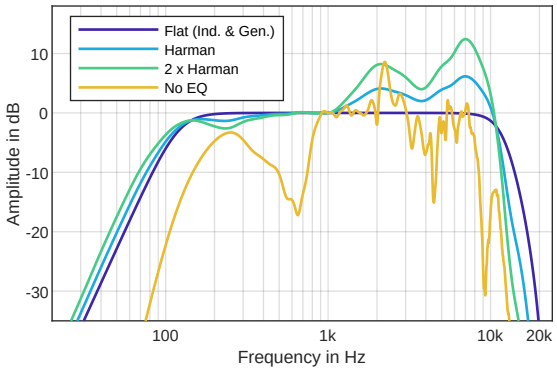

Fig. 4: Target HpTFs evaluated in the listening test.

As some of the evaluated target HpTFs were generated from generic measurements, it is relevant to explore how they may be perceived by different listeners. The case of the **generic flat** HpTF is illustrated in Fig. 5, which shows the statistics of HpTFs measured on 44 human subjects after applying HpEQ to the said HpTF. It can be observed that inter-subject spectral variance increases with frequency, in agreement with previous studies [19]. The differences in magnitude response across subjects even become comparable to the differences between **generic flat** and **Harman** targets. This high variance may lead to substantial differences in the effectively presented target HpTF during the listening test (e.g. for a given listener, the **Harman** target may sound "flatter" than the **generic flat** target). Furthermore, it seems that the median measurement deviates considerably from 0 dB, indicating that the chosen generic HpTF differs from the median frequency response of the measured population. This indicates that, although designed to acoustically represent the median human head and torso, KEMAR may not well represent the measured subjects in the case of the transfer function between the VR headset's built-in loudspeakers and the ears' entrance. A similar observation was made by Lindau and Brinkmann [4] for another head and torso simulator.

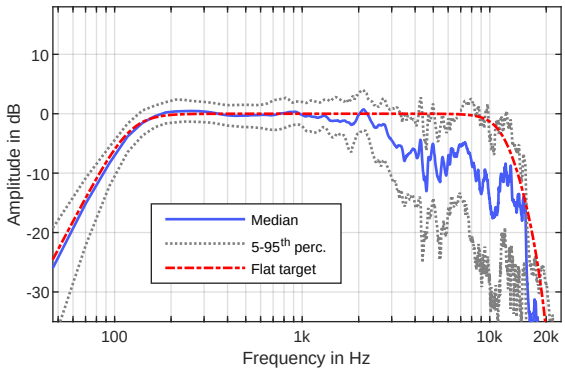

Fig. 5: Median, 5th, and 95th percentiles of the custom VR headset HpTF magnitude spectrum after applying **generic flat** HpEQ, measured at both ears of 44 subjects.

4 Listening test results

As described in section 2, a listening test was designed to evaluate VR headset users' preferred target HpTF for stereo and spatial audio contents. 21 listeners (13 male and 8 female) were recruited to participate in the listening test. 12 of them were Facebook employees (Redmond WA, USA) and 9 were external naive subjects, who were screened for normal audiometric hearing. Listeners' ages ranged from 18 to 55 years, with 12 individuals in the 25-34 years range.

In total, each of the 21 listeners performed 80 evaluations: 5 sound excerpts (one per evaluated target HpTF) in each trial, 2 repetitions of each trial, 4 audio materials and 2 content types (stereo and spatial audio). The 16 trials were divided in two blocks (one for each content type) and they were presented in a random order within each block. Listeners were encouraged to take a break to rest between blocks. The mean session time (not including breaks) across participants was 21.6 minutes, or 81 seconds per trial.

A post-screening method was used to exclude listeners whose ratings were not consistent across repeated trials more than once. Consistent ratings were measured by calculating the normalized cross-covariance between the ratings of two repeated trials. The post-screening criterion was to exclude any listener who displayed consistency that was below the tenth percentile of all analyzed data in at least two pairs of trials. 16 listeners (out of 21 listeners) passed the post-screening stage successfully.

Results are shown in Fig. 6 as the estimated marginal means and 95% confidence intervals of the preference ratings given to each evaluated target HpTF, for both audio content types.

The main question of interest is: which are listeners' preferred HpTFs for stereo and spatial audio contents? In order to answer this question we seek to show that there is a significant main effect of *target HpTF* on subjects' perceived audio quality, and that there is a significant interaction between *content type* and *target HpTF*.

Descriptive analysis shows that the **individual flat** target obtained the highest ratings for **spatial** content, while for **stereo** content the **Harman** target seems to obtain the highest ratings, although these should be confirmed by inferential analysis. The **no EQ** HpTF was rated the lowest for both content types, followed by **2 x Harman**.

A four-way repeated measures analysis of variance (rmANOVA), with factors *target HpTF*, *content type*, *material* and *trial*, and a significance level of 0.05, was used to analyze the main effects and the interactions between all main factors, as recommended by [17]. The rmANOVA with a Huynh-Feldt correction determined that subjects' perceived audio quality was statistically significantly affected by *target HpTF* ($F(2.7, 40.6) = 80.5$, $p < .001$), but not by *content type* ($p > .05$). In addition, the rmANOVA revealed a significant interaction between *target HpTF* and *content type* ($F(1.6, 24.2) = 13.3$, $p < .001$), which indicates that listeners' HpTF preference depends on whether the content is stereo or spatial audio. No significant main effect of *trial*, nor interaction with *target HpTF*, were found ($p > 0.05$), which, together with examination of the data, implies that subjects' ratings did not systematically change over trials.

Next, two separate three-way rmANOVA (factors: *target HpTF*, *material* and *trial*; significance level: 0.05) were conducted for every *content type*.

For **stereo** content, the rmANOVA with a Huynh-Feldt correction determined that *target HpTF* had a statistically significant effect on the perceived audio quality ($F(2.3, 33.9) = 38.5$, $p < .001$). Post hoc pairwise comparisons between all pairs of *target HpTF* revealed that the **Harman** target ratings were statistically significantly higher than for all other target HpTFs ($p < 0.05$). A significant effect of *material* ($F(2.3, 35.2) = 9.4$,

$p < .001$) was observed as well, while no significant effect of *trial* was observed.

For **spatial** content, the rmANOVA with a Huynh-Feldt correction also determined that *target HpTF* had a statistically significant effect on the perceived audio quality ($F(3.4, 51.2) = 95.2$, $p < .001$). In this case, post hoc pairwise comparison between all pairs of *target HpTF* revealed that the ratings for **individual flat** were statistically significantly higher than for all other target HpTFs ($p < 0.05$), which indicates that listeners' preferred target HpTF differs between **stereo** and **spatial** audio contents. A significant effect of *material* ($F(2.3, 35.2) = 9.4$, $p < .001$) was observed as well, while no significant effect of *trial* was observed.

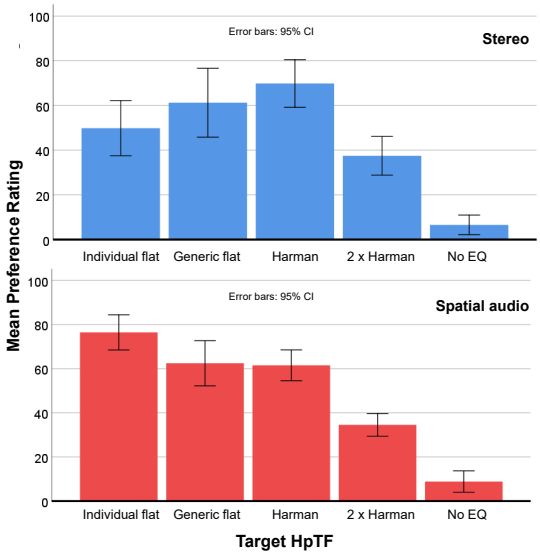

Fig. 6: Estimated marginal means and 95% confidence intervals of the preference ratings. Top: results for **stereo** content. Bottom: results for **spatial** audio content.

5 Discussion

Results show that the audio content had a significant effect on listeners' preference of target HpTF. For spatial audio content, **individual flat** was the preferred choice, which agrees with previous research [1, 2, 3]. For stereo content, on the other hand, **Harman** obtained higher ratings. This confirms the hypothesis that the audio content type should be taken into account when HpEQ is considered, if the aim is to optimize listener preference and individual HpEQ is available. For generic HpEQ, on the other hand, this effect was less evident, as **generic flat** and **Harman** targets did not show significant differences for spatial audio content. This result may be due to a number of reasons. First, it is possible that the presence of **individual flat** and **no EQ** targets, both of which produced extreme ratings towards either side of the scale, brought other targets closer together near the middle. Second, the perceptual difference between **generic flat** and **Harman** might actually be too small to produce significant differences in rating, with the current sample size. This effect may be emphasized by the high variance in measured HpTF across listeners, leading to each listener perceiving a different version of the intended target HpTF, as observed in the objective analysis. This would explain why when target curves deviate too much from the "ideal" target (e.g. **2 x Harman**), listeners generally give them lower ratings. To explore this hypothesis, future studies should test a new subset of target HpTFs which are perceptually similar to each other, thus affording finer detail in the perceptual comparison. Furthermore, the effect of reproduction bandwidth and headphone type should also be investigated in order to understand to what extent these results can be generalized.

6 Summary

This study explored the effect of audio content type on the preference of listeners with regard to headphone target frequency responses. It was shown that an individually calibrated flat response was the preferred choice for spatial audio (binaural) content, but not for stereo content, for which the Harman target was rated significantly higher. This explicitly confirms that content-dependent HpEQ would be beneficial for devices designed to reproduce both spatial and non-spatial audio, such as head-tracked headphones, VR headsets and AR glasses.

References

[1] 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.

[2] 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, doi:10.1121/1.397557.

[3] 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, doi:10.1121/1.417337.

[4] Lindau, A. and Brinkmann, F., "Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings," *Journal of the Audio Engineering Society*, 60(1/2), pp. 54–62, 2012.

[5] Engel, I., Alon, D. L., Robinson, P. W., and Mehra, R., "The Effect of Generic Headphone Compensation on Binaural Renderings," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, 2019.

[6] Møller, H., Jensen, C. B., Hammershøi, D., and Sørensen, M. F., "Design Criteria for Headphones," *Journal of the Audio Engineering Society*, 43(4), pp. 218–232, 1995.

[7] 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.

[8] Lorho, G., "Subjective Evaluation of Headphone Target Frequency Responses," in *Audio Engineering Society Convention 126*, Audio Engineering Society, 2009.

[9] Olive, S., Welti, T., and McMullin, E., "The Influence of Listeners' Experience, Age, and Culture on Headphone Sound Quality Preferences," in *Audio Engineering Society Convention 137*, Audio Engineering Society, 2014.

[10] Olive, S., Welti, T., and Khonsaripour, O., "A Statistical Model That Predicts Listeners' Preference Ratings of In-Ear Headphones: Part 1—Listening Test Results and Acoustic Measurements," in *Audio Engineering Society Convention 143*, Audio Engineering Society, 2017.

[11] Olive, S. and Welti, T., "Factors That Influence Listeners' Preferred Bass and Treble Levels in Headphones," in *139th Audio Engineering Society International Convention, AES 2015*, Audio Engineering Society, 2015.

[12] Olive, S., Welti, T., and Khonsaripour, O., "The Influence of Program Material on Sound Quality Ratings of In-Ear Headphones," in *Audio Engineering Society Convention 142*, Audio Engineering Society, 2017.

[13] Farina, A., "Advancements in Impulse Response Measurements by Sine Sweeps," in *Audio Engineering Society Convention 122*, Audio Engineering Society, 2007.

[14] Masiero, B. and Fels, J., "Perceptually Robust Headphone Equalization for Binaural Reproduction," in *Audio Engineering Society Convention 130*, pp. 1–7, Audio Engineering Society, 2011, doi:10.13140/2.1.1598.6882.

[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, doi:10.17743/jaes.2016.0030.

[16] Cabrera, D., Lee, D., Yadav, M., and Martens, W. L., "Decay Envelope Manipulation of Room Impulse Responses: Techniques for Auralization and Sonification," in *Acoustics 2011*, p. 5, Gold Coast, Australia, 2011.

[17] ITU-R, "BS.1534: Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems," 2015.

[18] ITU-T, "P.57: Artificial Ears," 2011.

[19] 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.