Virtual Headphone testing for Spatial Audio

Hugh O'Dwyer[1], Marcin Gorzel[2], Luke Ferguson[1], Enda Bates[1] and Francis M. Boland[1,2]

[1] Trinity College Dublin, Dublin, Ireland.

[2]Google, Dublin, Ireland.

Correspondence should be addressed to odwyerh@tcd.ie

September 23rd 2016

Abstract

With recent advances in Virtual Reality (VR) systems and an increased interest in the medium for gaming and 360-degree cinematic experiences, there is a need to establish a suitable method for comparing audio systems in VR. Typically, VR systems incorporate headphones for playback of dynamic, spatial audio. Several methods of performing comparative headphone testing exist today with virtual methods being preferred to the conventional method, which involves physically switching and comparing headphones. Comparative listening tests on multiple headphones can be challenging to conduct in a controlled environment using a double-blind methodology and many potential biases can influence a listener's preference including brand, price and design. Virtual headphone testing has been introduced to minimise these factors in order to concentrate only on the sound quality. This paper compares two methods of virtual headphone testing for their ability to quantify the quality of spatial audio over low cost headphones. The first method uses recordings made from each of the headphones under test using a dummy head microphone. The second and more widely used method uses the transfer functions of each headphone to generate stimuli using convolution. Using the above methods, 4 pairs of headphones were compared with respect to their to their overall perceived quality as well as spatial impression. Although there was a good correlation between the results of the different tests, more detailed statistical analysis revealed significant differences, particularly with regards to evaluation of spatial impression.

1 Introduction

The delivery of spatial audio with a high degree of sound quality and accurate localisation is essential for creating a sense of immersion in a virtual environment. Typically, spatial audio in virtual environments is delivered via headphones. Other means include using speaker arrays although these methods can be problematic due to user's natural tendencies to move and adjust ones position. Thus, it is important to develop an accurate method of comparing the sound quality and localisation of spatial audio across a range of headphones, in an attempt to quantify what properties are most significant in creating a sense of immersion in a virtual environment. Virtually immersive games and visual experiences are becoming increasingly prevalent as leading tech companies such as Facebook, HTC, Sony and Google are launching high-end virtual content and the technology to present it, taking the Oculus Rift, HTC Vive, Playstation VR, and Youtube's new 3D video presentation as examples of this.

With the standard of the visual component of virtual experiences increasing so rapidly it is important that the standard of accompanying spatial audio improves at a similar rate. Establishing a suitable method of audio playback is essential to this and determining the most accurate method of headphone testing is fundamental in this regard as headphones are most often the preferred choice for audio playback in virtual environments.

This paper describes a testing process conducted to compare two methods of virtual headphone testing. Section 2 discusses some of the relevant work that has been done in relation to virtual headphone testing and spatial audio. In section 3, two methods of virtual headphone testing, 'filtered' and 'recorded' are described as well as how we performed them for our own comparative study. A statistical analysis of our results is described in Section 4 while Section 5 concludes the paper with a summary of our findings.

2 Related Work

In recent research literature, three test strategies have been employed for the comparative testing of headphones. The first of which is a conventional swap and compare approach, which can result in multiple biases due to factors such as headphone design, fit and branding [1]. A popular alternative to the conventional approach to headphone testing is the virtual test approach which aims to remove any bias by having the subject use just one set of headphones throughout the testing procedure [1,2,3]. These monitoring headphones are of a high quality and have a relatively flat frequency response and it is assumed that any coloration caused by these will not have a significant effect on the listener's preference of the audio. Typically, an inverse filter is applied to test stimuli to equalize the affect of these headphones [1,2,3].

In the first virtual method, listeners are presented with a series of test samples that have been produced by another set of headphones and recorded onto a binaural head microphone [1]. This is the 'recorded' test strategy presented by Hirvonen et al [1]. Based on the preference ratings of 21 subjects, the study showed that results attained for the 'recorded' method and the conventional swap method were similar. However, statistical analysis showed significant differences beyond the confidence intervals meaning the virtual method could not be validated based on these results. That study was also limited to speech signals and was conducted at a sampling frequency of 10kHz, which would have major impact on the perception of higher frequency components [1].

More recently, a new virtual headphone testing methodology was proposed by Olive et al [2]. This method presents the listener with virtualized versions of multiple headphones through a single reference headphone that is equalized to simulate the linear magnitude response of the different headphones. In this study, the Headphone Related Transfer Functions (HpTF) of several headphones were measured using an ear and cheek simulator microphone. Although it has been shown that there is a wide degree of variability in the characterization of the headphone transfer function [4], multiple reseatings of the headphones and an averaging process are set in place to account for this. These transfer functions are then convolved with a set of stimuli to simulate the sound of the headphones and listeners are asked to rate their preference of virtualized headphone sounds as well as rating the same headphones using the conventional method. A substantial set of listening tests examining preference and spectral balance showed a good correlation between the two methods when examining overall headphone preference (correlation coefficient r = 0.85) [2]. Following on from this study, further research has been conducted to reduce the error in estimating these HpTFs [5]. The Princeton Headphone Open Archive (PHOnA), a database of HpTFs, has since been established and shared with the public in an attempt to optimize equalization algorithms to provide more universal solutions to perceptually transparent headphone reproduction [6].

A study published in April 2016 investigated the influence that the perceived quality of consumer headphones has on the spatial impression of the audio [3]. Subjects were presented with a virtual headphone comparison method similar to the one explored in [2] and were asked questions based on both the timbral quality and the spatial impression of the headphones. The study concluded that there is a strong correlation in the perceived quality and the spatial impression of the headphones under test in the study. The virtual testing conducted in this study was presented to subjects using a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) type test [8]. This required subjects to rate the headphones under test in relation to a high quality reference headphone as well as a low quality 'anchor' rendering of the high quality reference [8]. For questions regarding perceived quality, a low passed render of the reference was used as the anchor whereas a monaural render of the reference was used for questions regarding spatial impression.

Previous studies addressing the differences between headphone comparison methods have only addressed one virtual testing method with relation to physical headphone testing [1,2]. In this study we aim to compare two virtual methods of headphone testing to determine whether the results they provide are significantly different. In addition to comparing these two methods we also wish to examine the correlation between perceived audio quality and spatial impression and how this may differ for the two methods. As it was shown in [3] that these two measures are highly correlated when using a 'filtered' virtual test it would be interesting to observe whether this may be the case while using the 'recorded' virtual method.

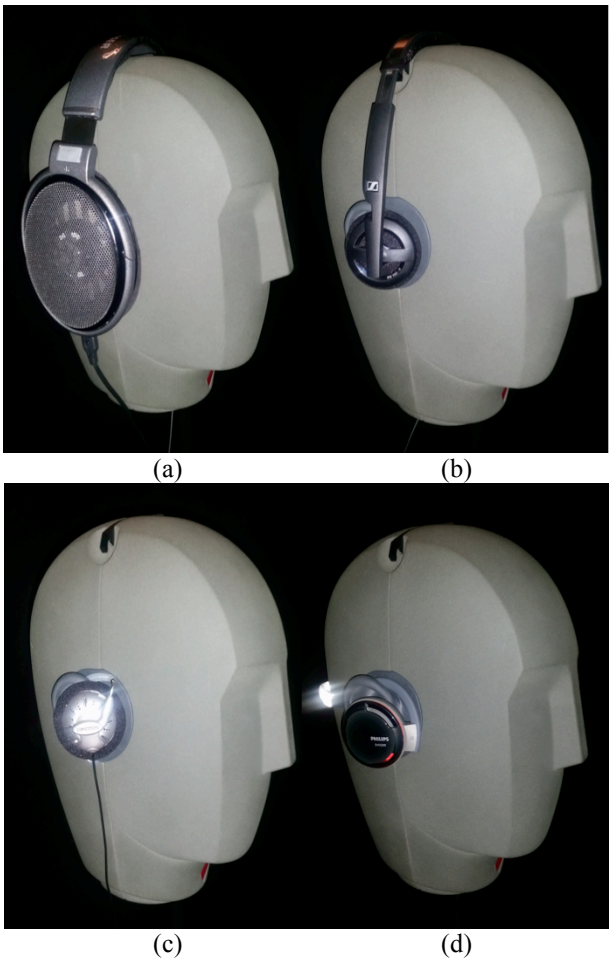
3 Methodology

In this study, two methods of virtual headphone testing as described in [1 & 2] are compared to determine whether their results are significantly similar. As previous attempts to validate the testing method described in [1] have been successful, it would be beneficial to determine whether the results yielded from this type of testing can be comparable to those attained from a method, which has been validated. MUSHRA

type testing as described in [3] was used to validate and compare both methods. These methods are described in further detail throughout this section.

3.1. Headphones

Four headphones were compared in this study. Sennhesier HD650 headphones were chosen as a high quality reference while the other three headphones were chosen based on their low cost and unique designs. These can be seen in the table 1 below.



(a) (b) (c) (d)

Figure (1): The 4 headphones under test mounted on the Neumann KU-100 microphone, (a) Sennheiser HD650, (b) Sennheiser PX-100-II, (c) KOSS KSC75, (d) Philips SHS5200.

For both virtual tests, Beyerdynamic DT-770 PRO headphones were used to playback the virtual renderings of each of the headphones under test. An inverse filter was applied to each stimulus to account for the effect of these playback headphones.

Table (1): The headphones examined in this study:

Manufacturer	**Model**	**Price, €**	**Type**
Sennheiser	HD 650	315*	Over-Ear, Open, Reference
Sennheiser	PX-100-II	53**	On-Ear, Open
Koss	KSC 75	21**	On-Ear
Philips	SH 5200	23.50**	On-Ear

**Price from thomann.de as of September 2016*
***Price from Amazon.co.uk as of September 2016*

Figure (2): The headphones used in both virtual tests to playback the virtual stimuli, Beyerdynamics DT-770-Pro.

3.2. Stimuli

Table (2): The stimuli used in this study:

Stimulus	**Description**	**Length, seconds**
Mark Ronson ft. Bruno Mars - Uptown Funk	Up-tempo pop song with high quality production and a wide sound stage and responsive bass	~26
Nature Scene	A binaural recording of a thunder storm with moving sound sources	~30
Binaural Shaker Recording	A recording of a shaker moving about a binaural head microphone	~20
Surround Sound Speaker Test	A binaural rendering of 5.1 speaker surround test with 5 separate source locations (Front Right, Front Left, Centre, Back Left & Back Right)	~10

Four stimuli were selected for use in this study which represented a wide range of spatial audio. These stimuli differed in length with the shortest being ~10 seconds long and the longest being ~30 seconds long. A

description of each of these Stimuli can be found in table 2.

3.3. Virtual Testing

Two methods of virtually testing were employed to evaluate the perceived quality and spatial impression of each the headphones. In these tests, participants listened to four audio samples as well as a low quality anchor. Each passage was simulated to sound as it would through each of the four headphones under test. Two low quality anchors were implemented in these tests. A lowpassed rendering of the reference headphone signal to 3.5kHz was used as the anchor for questions regarding timbral quality and a mono rendering on the reference signal was used for questions regarding spatial quality. The anchor signals are included as per the MUSHRA test guidelines to ensure there is both a high quality and low quality reference for subjects to compare the test signals to [8].

Participants listen to these simulations using a single set of monitor headphones (Beyerdynamic DT-770 PRO, a closed back headphone). These headphone simulations were created in separate ways for each of the two tests. The test samples used in the first virtual test were simulated by convolving the frequency response of each of the headphones with the stimuli. For the second virtual test, each of the four stimuli were recorded over the four headphones under test using a dummy head microphone, the Neumann KU-100. These recordings were then used as the test samples in the virtual test. Both sets of virtual test samples were equalized to remove the effect of the monitor headphones and were normalized to -3dB to ensure the same volume was heard across the four headphone simulations.

These perceptual tests were designed according to the International Telecommunications Union, Radio communication sector (ITU-R) recommendation BS.1534-1 [8]. This recommendation describes the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) perceptual test which is intended for spatial audio psychoacoustic tests. This test uses a reference sound to comparatively evaluate the quality of several other lower quality sounds. Using a continuous scale from 0 to 100, samples are graded by participants on their relative quality compared to the reference sample. Subjects were asked eight individual questions, each time evaluating either the timbral or spatial quality of the four test passages in comparison to that of the reference.

The testing consisted of a MUSHRA test which was split into two parts, the first using filtered stimuli made using HpTFs and the second using recordings of the stimuli made over headphones onto a dummy head microphone. Both parts of the test consisted of eight questions, four comparing the timbral quality for each of the stimuli and the other four comparing the spatial quality of the same stimuli. There were four headphone signals under comparison for each question as well as a low quality anchor signal. In total, subjects attended to 80 stimuli (two tests x eight questions x five stimuli). On average testing took approximately 40 minutes to complete.

3.4. Measuring the Frequency Response of each Headphone

The frequency response of each headphone was measured using the Neumann KU-100 dummy head microphone. This was achieved by playing a swept sinusoid at 44.1kHz/24bit through each set of headphones mounted on the dummy headphone. The recordings were made in an acoustically treated room with a background noise level of 27.2dBA and a RT60 of 0.13 seconds at 1kHz. The recordings made of the swept sine tone were deconvolved with the source signal using Voxengo's Deconvolver [9]. Impulse responses from each headphone were thus obtained and were transformed in the frequency domain using a fast Fourier transform to give us the resulting frequency response. These measurements were repeated 5 times and averaged to account for the variability in the characterization of the headphone transfer function due to variations in the coupling with the ear [4]. The resulting HpTFs can be seen in figure (3). These transfer functions were then convolved with the original stimuli to simulate the sounds of the stimuli through each headphone.

3.5. Recording stimuli for test 2

Similarly to capturing the frequency response of each of the headphones, the dummy head recordings were performed in the same room under the same conditions. Input levels for both the left and right channels of the KU-100 were equalized and each of the four stimuli were recorded though each of the four headphones under test. Each stimulus had a sampling rate of 44.1kHz and had a bit rate of 32 bit floating point. The Beyerdynamic headphone inverse filter was applied to the recordings, which were then normalized to -3dB before being used implemented in the test.

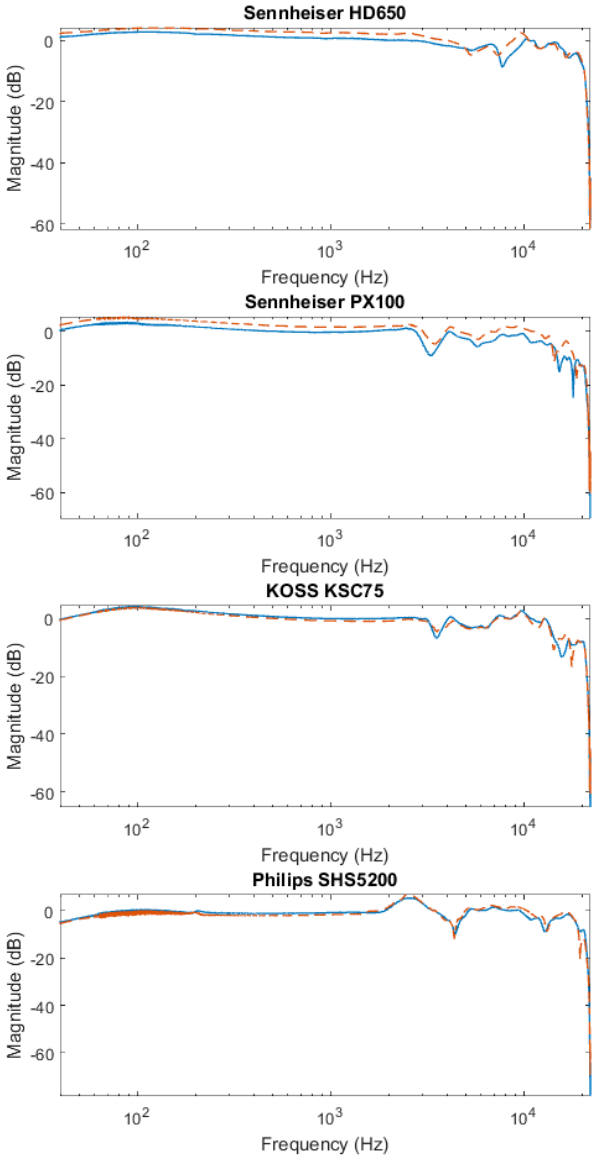


Figure (3): Frequency Response of the headphones used in the study. Solid curves, left channel; dashed curve, right channel.

4 Experimental Results

A total of 10 subjects took part in these MUSHRA tests with none being disqualified for results that did not satisfy the criteria of the ITU-R recommendation. The average results and standard errors for each question have been included in the plots below.

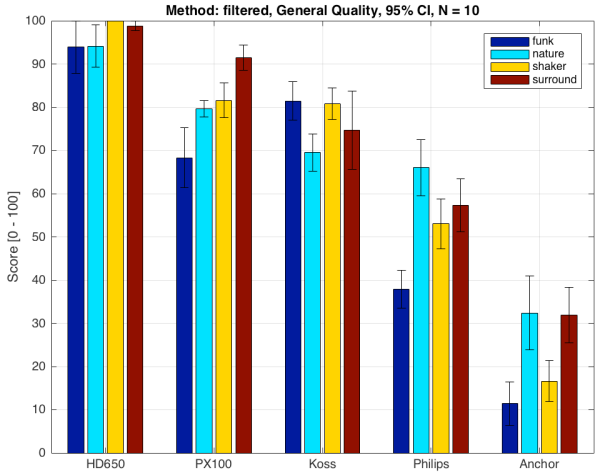


Figure (4): Results for the questions regarding general quality using the filtering method.

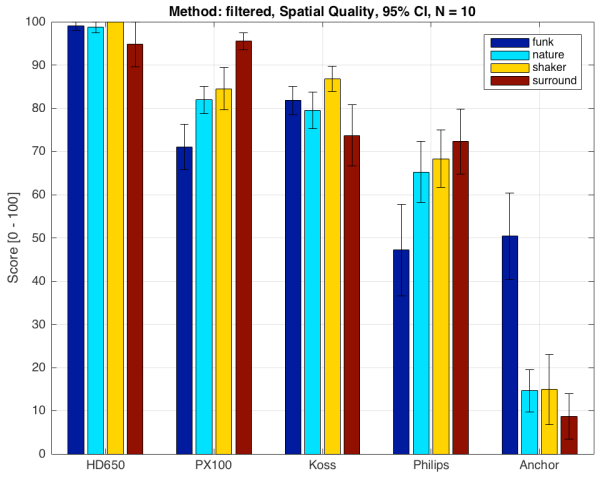


Figure (5): Results for the questions regarding spatial quality using the filtering method.

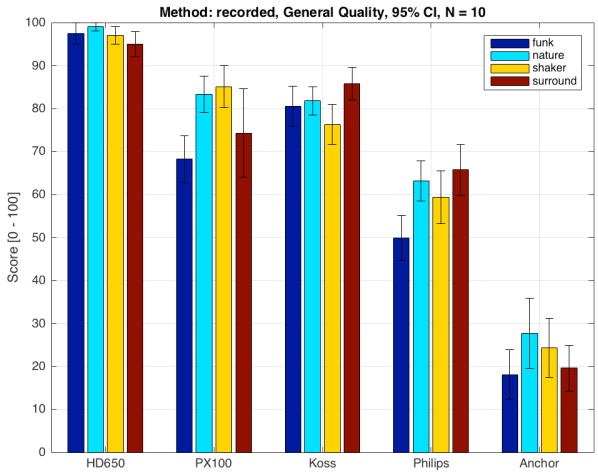


Figure (6): Results for the questions regarding general quality using the recorded method.

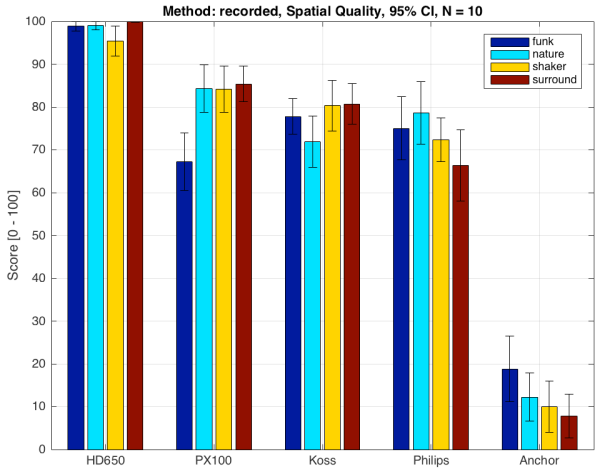


Figure (7): Results for the questions regarding spatial quality using the recorded method.

As can be seen in the plots above, there is a similar ranking of the four headphones for both the filtered and recorded methods as well as for the questions relating to perceived quality and spatial impression. High correlation between all four test results have been identified as illustrated in tables 3 & 4.

To further investigate the possible differences in the results, 1-way ANOVA testing was performed for each method and for both questions regarding general and spatial quality preference. Results across all the test samples were included in the comparison. Pairwise comparisons of all the headphone pairs were conducted with *p-values* being computed using Tukey's HSD method in order to determine whether the differences are statistically significant. Matrices of these values can be found in the appendix of this paper (tables 5 through 8).

The results again indicate that both test methods find differences between the headphones in a very similar way. Insignificant differences between the rating of the Koss headphones and the Sennheiser PX 100 headphones are consistent for both general and spatial quality for the filtered stimuli and for the questions relating to general quality for the 'recorded' stimuli. The only exception is that the Spatial Quality test in which the 'recorded' method fails to find differences between 3 lower-grade models. That might suggest that there is a difference between the results attained for the two tests and therefore they are not comparable. However, more data would be required to confirm that finding.

Another interesting point is that when rating the spatial quality of the monaural anchor for one of the filtered stimuli, a relatively high score was achieved compared to the other monaural anchors and compared to the monaural anchor in the recorded stimuli. This result is most likely due to the high quality of this particular music sample and how it was most likely mixed to still provide a balanced, high quality track even when downmixed to mono.

5 Conclusion

This study aimed to demonstrate whether two different methods of virtual headphone testing produced similar results. We have presented some evidence to suggest that these two methods which although produce similar results are significantly different. Previous studies have concluded that using HpTFs to generate stimuli for virtual headphone testing produces similar results to that of physical headphone testing [2,3]. Other studies suggest that recording stimuli on to binaural microphones does not produce results that are significantly consistent with physical headphone testing [1]. It has also been shown that for questions regarding perceived audio quality and questions regarding spatial impression, results are significantly similar in a MUSHRA type test [3].

For the virtual test using HpTFs we found that there was a strong correlation between the rating of headphones for questions regarding perceived quality and spatial impression. The results gathered for the Sennhesier PX 100 and Koss headphones were significantly indifferent while the reference headphone performed the best and Philips headphones performed the worst after the anchor stimulus. This was also the case when comparing perceived audio quality for the recorded method.

However, for questions regarding spatial quality for the recorded method, the Philips headphones scored similarly to the Koss and PX 100 headphones. This demonstrates as inconsistency for the recorded method that would suggest that the filtered method is a significantly more reliable method when performing headphone comparison testing.

References

[1] Hirvonen, Toni, et al. "Listening test methodology for headphone evaluation."*Audio Engineering Society Convention 114*. Audio Engineering Society, 2003.

[2] Olive, Sean E., Todd Welti, and Elisabeth McMullin. "A virtual headphone listening test methodology." *Audio Engineering Society Conference: 51st International Conference: Loudspeakers and Headphones*. Audio Engineering Society, 2013.

[3] Gutierrez-Parera, Pablo, and Jose J. Lopez. "Influence of the Quality of Consumer Headphones in the Perception of Spatial Audio." *Applied Sciences*6.4 (2016): 117.

[4] Kulkarni, Abhijit, and H. Steven Colburn. "Variability in the characterization of the headphone transfer-function." *The Journal of the Acoustical Society of America* 107.2 (2000): 1071-1074.

[5] Boren, Braxton, et al. "Coloration metrics for headphone equalization." (2015).

[6] Boren, Braxton B., et al. "PHOnA: a public dataset of measured headphone transfer functions." *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.

[7] Bouchard, Martin, Scott G. Norcross, and Gilbert A. Soulodre. "Inverse filtering design using a minimal-phase target function from regularization." *Audio Engineering Society Convention 121*. Audio Engineering Society, 2006.

[8] International Telecommunications Union, ITU-R Recommendation BS.1534-1, 2001-2003.

[9] http://www.voxengo.com/product/deconvolver/

Appendix

Table 3: Correlation coefficients (R)

Test	Filtered-general	Filtered-spatial	Recorded-general	Recorded-spatial
Filtered-general	1	0.902	0.958	0.906
Filtered-spatial	0.902	1	0.926	0.931
Recorded-general	0.958	0.926	1	0.947
Recorded-spatial	0.906	0.931	0.947	1

Table 4: Correlation *p-values*:

Test	Filtered-general	Filtered-spatial	Recorded-general	Recorded-spatial
Filtered-general	-	<0.001	<0.001	<0.001
Filtered-spatial	<0.001	-	<0.001	<0.001
Recorded-general	<0.001	<0.001	-	<0.001
Recorded-spatial	<0.001	<0.001	<0.001	-

Tables 5-8 below visualize pairwise comparisons of all the headphone pairs and use *p-values* computed using Tukey's HSD method in order to determine whether the differences are statistically significant, *p-values* larger than 0.05 (in bold) indicate that the difference is not significant.

Table 5: General quality preference. Test method: filtered.

Headphones	HD650	PX100	Koss	Philips	Anchor
HD650	-	<0.001	<0.001	<0.001	<0.001
PX100	<0.001	-	**0.89**	<0.001	<0.001
Koss	<0.001	**0.89**	-	<0.001	<0.001
Philips	<0.001	<0.001	<0.001	-	<0.001
Anchor	<0.001	<0.001	<0.001	<0.001	-

Table 6: Spatial quality preference. Test method: filtered.

Headphones	HD650	PX100	Koss	Philips	Anchor
HD650	-	0.008	<0.001	<0.001	<0.001
PX100	0.008	-	**0.97**	<0.001	<0.001
Koss	<0.001	**0.97**	-	0.001	<0.001
Philips	<0.001	<0.001	0.001	-	<0.001
Anchor	<0.001	<0.001	<0.001	<0.001	-

Table 7: General quality preference. Test method: recorded.

Headphones	HD650	PX100	Koss	Philips	Anchor
HD650	-	<0.001	<0.001	<0.001	<0.001
PX100	<0.001	-	**0.9**	<0.001	<0.001
Koss	<0.001	**0.9**	-	<0.001	<0.001
Philips	<0.001	<0.001	<0.001	-	<0.001
Anchor	<0.001	<0.001	<0.001	<0.001	-

Table 8: Spatial quality preference. Test method: recorded.

Headphones	HD650	PX100	Koss	Philips	Anchor
HD650	-	<0.001	<0.001	<0.001	<0.001
PX100	<0.001	-	**0.96**	**0.34**	<0.001
Koss	<0.001	**0.96**	-	**0.76**	<0.001
Philips	<0.001	**0.34**	**0.76**	-	<0.001
Anchor	<0.001	<0.001	<0.001	<0.001	-