PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Loudness in different rooms versus headphone reproduction: Is there a mismatch even after careful equalization?

Michael Kohnen[(1)], Florian Denk[(2)], Josep Llorca-Bofi[(1)], Michael Vorländer[(1)], Birger Kollmeier[(2)]

[(1)]Institute of Technical Acoustics, RWTH Aachen University, Germany, michael.kohnen@akustik.rwth-aachen.de
[(2)]Medizinische Physik & Cluster of Excellence Hearing4All, Universität Oldenburg, Germany, florian.denk@uni-oldenburg.de

Abstract

Even though our expectation is that sound reproduced by loudspeakers in a room should exert the same perceived loudness if reproduced by headphones with the same sound pressure level at the eardrum, findings from literature (e.g., Munson 1952[1], Fastl et al. 1985[2]) indicate a mismatch: Equal sound pressure levels at the eardrum do not guarantee an equally perceived loudness. The aim of this study is to characterize this mismatch as a function of room acoustics and headphone type. Subjects balanced the perceived loudness between presentations of an identical stimulus from a loudspeaker and a headphone for three different rooms (anechoic, office room, reverberation chamber) and two different headphones (Sennheiser HD650 and Beyerdynamic DT770 PRO). Additionally, a comprehensive acoustic characterization of the headphones with different measurement devices was conducted. The results confirm a loudness mismatch dependency on the room type, indicating that room acoustics biases our loudness judgment. This is important for sound reproduction systems in rooms and the limited fidelity of room simulations even with carefully equalized headphones. The consequences of our findings and necessary steps to pin down the origin of the observed mismatch will be discussed.

Keywords: loudness, room acoustics, sound reproduction

1 INTRODUCTION

Audio reproduction for plausible or even authentic scene representation is demanded in modern 3-D virtual and augmented reality systems. Especially in augmented reality applications, the user compares the perceived sound image to an environment that is real and familiar from everyday life, lowering accepted tolerances for mismatches. This investigation focuses on the perception of correct loudness in headphone reproduction as one part of these mismatches. Loudness encodes information about the distance to a source and its geometrical attributes. For room auralizations it relates to the applied materials in the room in terms of reflection energy. If a certain force is the cause of sound generation, loudness relates to the amount of that force. Finally, for virtual presentations of humans, changes in loudness shifts the perception of the intended expression and respectively mood to be emulated. Wrong loudness in realistic virtual and augmented reality therefore distracts the user from a full immersion into the scene. And finally, long term listening over headphones using higher gains to compensate mismatches can be more hazardous than listening over loudspeaker.

Having a calculated reference SPL at the eardrum for the headphone reproduction, it might still differ in its perceived loudness compared to sound generated over the loudspeakers that results in the same measurable eardrum SPL. This mismatch (also known as the *missing 6 dB* [1]) is known for over 70 years [3] now in the area of loudness balancing. In blind comparison of headphone and free-field listening and when the same waveform is reproduced at the ears using individual BRIR convolution and headphone compensation, no mismatch occurrs [4, 5]. However, in this study we assess the loudness mismatch in "classic" headphone listening without BRIR convolution and headphone compensation, and in a non-blind comparison to loudspeaker presentation. As the mismatch is investigated in different set-ups by different groups, one additional question is: How reproducible are the fundamental measurements on which such a study relies? This work investigates the differences due to different measurement devices (namely a headphone test fixture and KEMAR head and torso simulator). It first investigates the headphone measurements as a basis for a listening test done using these headphones. In the end

results are analyzed, discussed and concluded. As part of the listening test, the influence of binaural correlation was investigated. Evaluation for this can be found in [6].

2 HEADPHONE MEASUREMENTS

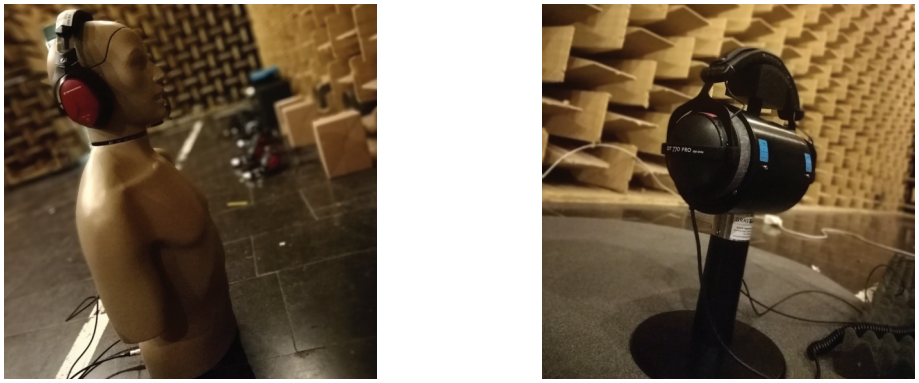

Figure 1. Picture of the KEMAR with Sennheiser HDA300 (left) and headphone test fixture with Beyerdynamic DT770 PRO (right).

To understand the influence of headphones, three different headphone types were measured on two different devices with different ear simulators. Headphones measured were the Sennheiser HDA300 and the two headphones used for the listening test, the Sennheiser HD650 and the Beyerdynamic DT770 PRO. The first one is a fully closed reference headphone used for audiometric measurements and subject screening. The second one is an open headphone known for a flat frequency response. The last one is characterized by its closed design. The

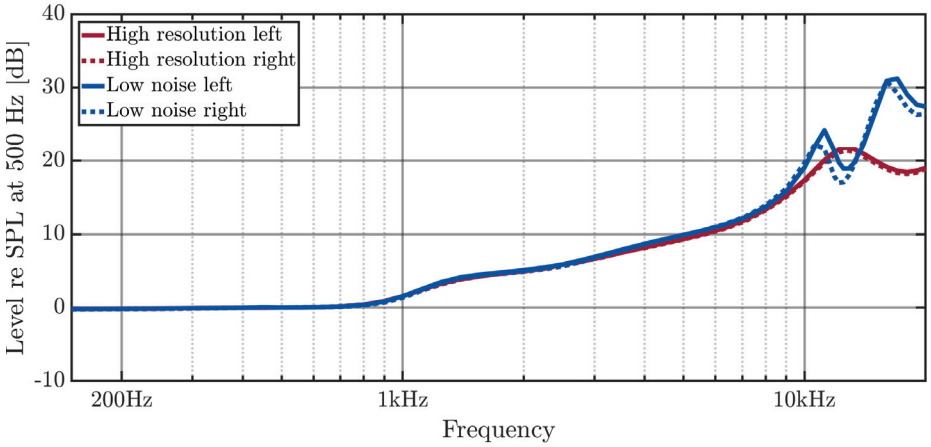

Figure 2. Frequency responses of the high-resolution ear simulator as used in the headphone test fixture and the low-noise ear simulator as used in the KEMAR artifical head and torso. The curves are divided by the sound pressure level at a frequency of 500 Hz and show an identical behavior ($\pm$0.65 dB) for frequencies below 9 kHz.

measurements for the eardrum sound pressure level were done on the one hand using a KEMAR 45BB-12 head and torso simulator with anthropometric pinna and a low-noise ear simulator and on the other hand using the

GRAS headphone test fixture 45CA-9 with high-resolution ear simulator and the same anthropometric pinna. The main difference between the low-noise ear simulator and the high-resolution one is the frequency response above 10 kHz. In fig. 2 the frequency responses are shown for both ear simulators divided by their sound pressure level at 500 Hz. The high-resolution ear simulator has a suppressed resonance at 13.5 kHz (compared to a standard IEC-711 ear simulator) while the low-noise has a double sloped resonance.

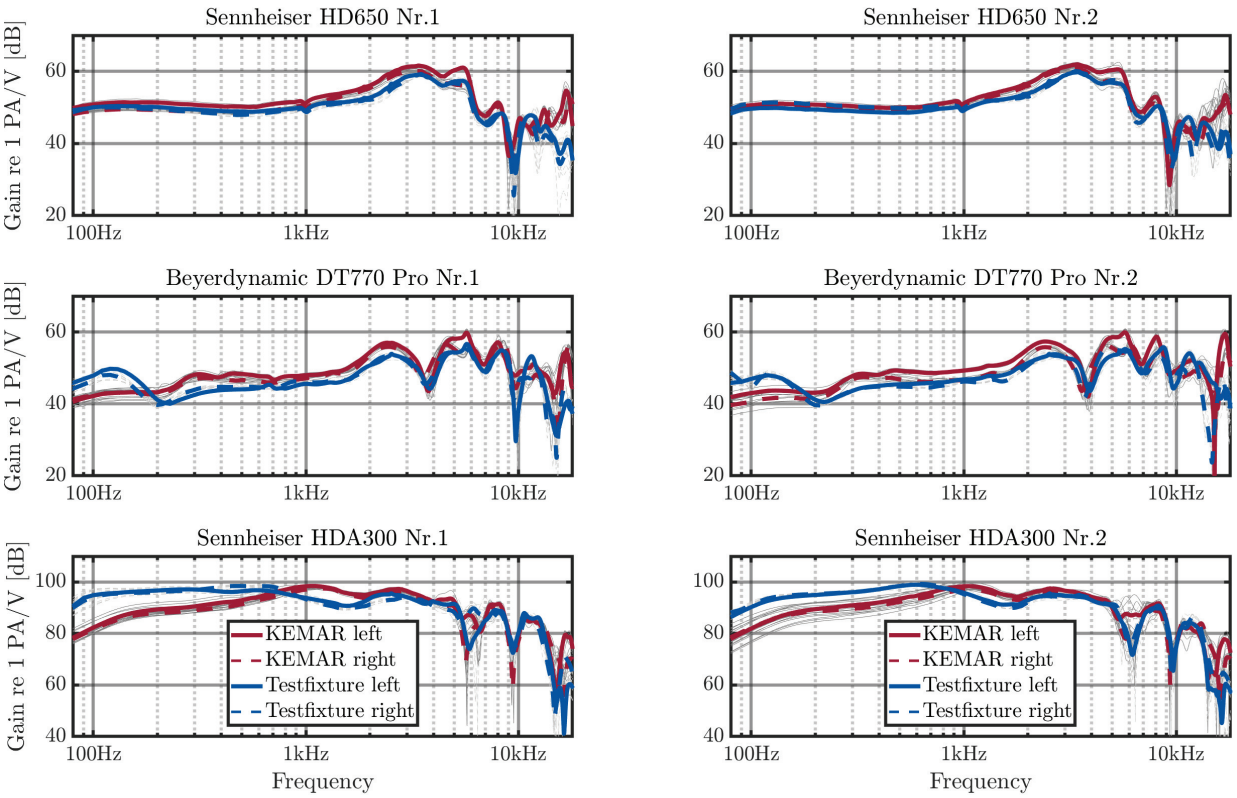

Figure 3. Measurements of the headphone transfer function for three different headphone types: Sennheiser HDA300 and HD650 as well as Beyerdynamic DT770 PRO. Gain describes the factor from input voltage to measured sound pressure at the eardrum. The gray lines indicate single measurements. On each measurement device the measurement was repeated eight times, each time the headphones were re-fitted on the measurement device. The mean values of the measurements are plotted as shown in the legend: red indicates the KEMAR measurement, blue the headphone test fixture measurement and solid lines the left, dotted lines the right ear.

For each measurement, the headphone was put on and off the measurement device to achieve different fittings. Eight repetitions (as suggested in [7]) were made using a exponential sweep of 2.7 seconds. The transfer functions for the HDA300 can be found at the bottom of fig. 3. Four different HDA300 were measured but only two are plotted exemplary as the results do not differ much. The high absolute values are due to a very small impedance of the headphones. The results show a good match for left and right channel for each measurement device. At frequencies above 10 kHz the results differ as can be expected due to the different ear simulator types. Two main differences between the devices can be observed for all headphones of the type HDA300: For low frequencies the measured level drops for the KEMAR and between about 1 kHz to 2 kHz the behavior is inverse. The low-frequency behavior can be explained due to different coupling. The headphone

test fixture as a plane surface around the ears where coupling is assumed to be best, as long as circum-aural headphones are used. The same effect can be seen for the closed DT770. As the HD650 are open headphones there is no such effect as can be seen in the upper part of fig. 3. The difference in the mid-frequency region is yet unsolved but might relate to the fact that the HDA300 do press on the ear and not fully surround it and lead to resonance due to a specific volume coupled to a leakage. The middle row of fig. 3 shows the frequency responses for the DT770. Here, only two devices were measured. Again a low-frequency difference can be observed and a slight offset up to 10 kHz. Above 10 kHz again both ear simulators result different levels but also seem to be less consistent between the two devices and measurements (e.g. notch depth at 15 kHz). Also misalignment between left an right ear is minimal. In general both measurement devices seem to result the same frequency response.

The HD650 is an open headphone and the measured frequency responses are shown in the upper row of fig. 3. Again, four different headphones were measured and two are shown exemplary. Left and right ear are aligned as well as the differences between the measurement devices is smaller than the differences between different measurements. The open design eliminates differences due to coupling. The only consistent difference is a slightly higher, flat peak around 5 to 6 kHz.

3 METHODS

To pin down the mismatch in perceived loudness between headphone and loudspeaker reproduction, a listening test was conducted. During the listening test, participants had to rate whether they perceived the loudspeaker (a GENELEC 8020C) or the headphone reproduction as louder, and the presentation level of the headphone was adapted to equal loudness. As possible influence, different rooms as well as two different types of headphones were investigated. Additionally different stimuli were used to investigate the frequency dependence of the perception.

3.1 Listening test design

The listening test was designed as a 1-up-1-down adaptive 2-AFC procedure using the AFC Toolbox [8, 9]. The step-size was adapted to 10, 5 and 1 dB, with 10 dB being the step size until the first reversal, and 1 dB in the measurement phase. Each stimuli was repeated three times and each block was done two times in each room, once using the HD650 headphone and once with the DT770. The stimuli order, room order and headphone order was randomized. As input device a pedal switch was used to start the stimulus and to respond whether the loudspeaker or the headphone reproduction was perceived as louder. The pedals were labeled with 'loudspeaker louder', 'headphone louder' and 'next'. Using the pedals the participants had both hands free to hold and put on and off the headphones. To reduce time and effort for putting on and off the headphones the order of stimulus presentation was loudspeaker vs. headphone and then headphone vs. loudspeaker (and from the beginning). The stimuli were each calibrated using an omni-directional free-field microphone (B&K 4190) to a level of 70 dB (re $20\mu Pa$). The level was calculated excluding the reverberant tail of the stimuli recording.

3.2 Stimuli

To investigate the frequency dependency of the mismatch, different one third octave band noise (tbn) stimuli were presented with mid-band frequencies of 250, 1000 and 4000 Hz using diotic playback for the headphones. The 250 Hz stimulus represents low-frequency features with higher binaural interaction and a mismatch should definitely be observable according to literature. 1000 Hz is the beginning of the transition from perception of phase cues to perception of level cues while 4000 kHz is above this transition and no phase interaction influences the perception. Additionally, the choice of a 1000 Hz tbn plus-minus 2 octaves make the findings more comparable to other studies. As narrow-band excitation in different regions of binaural interaction might neglect interaction of these effects, a unified excitation noise [10] with the same energy in the 17 auditory bands around 1 kHz was used that covers the whole range of binaural interaction. The overall length of each stimulus was one second inclunding 20 ms hanning ramps.

3.3 Rooms

Three different rooms were used for the listening test. The rooms are located at the Institute of Technical Acoustics in Aachen. The first one is a hemi-anechoic chamber, the second one an empty office room with a kitchenette and the third one a reverberation chamber. All three rooms can be seen in fig. 4.

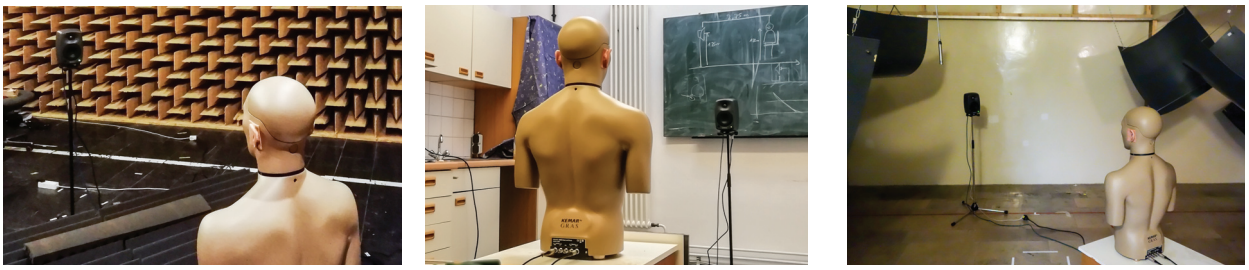

Figure 4. The pictures show the three different rooms in which the listening test was conducted. From left to right: Hemi-anechoic chamber, office room and reverberation chamber.

While the choice for the hemi-anechoic room and reverberation chamber as extrema was straightforward, the selection of the office room underlies some preceding thoughts. The room should not be to dry, so that a difference to the anechoic chamber can be observed, yet it should not have a too long reverberation time or distinct reflections such as flutter echoes or focus points that lead to an unnatural room perception. Therefore, an office room that is almost empty was chosen. It has a reverberation time of 0.6 seconds (T20, averaged between 0.5 and 1 kHz). Additionally, the room with kitchenette was chosen because it features some diffusion and therefore avoids symmetric effects of the wall. Room impulse response measurements were done with a Genelec 8020C loudspeaker at a distance of 2.25 meters (as in the listening test). In the hemi-anechoic chamber the floor was covered with absorbers to suppress floor reflections. The room is mainly characterized by its volume of about 300 m^3 and its lower cut-off frequency at 100 Hz due to the length of the wedges. The reverberation time for the reverberation chamber is 4.3 seconds.

3.4 Results

Eight test subjects (6 male, 2 female) participated in the listening test (except for the reverberation chamber where one person is missing due to logistical reasons). The reference sound pressure levels were measured with the KEMAR and low-noise ear simulators mounted. For loudspeaker reproduction, the KEMAR was positioned at the point where the participants were seated. Each calibrated stimulus was played back and recorded five times and averaged. The stimuli used in headphone reproduction were measured with the same KEMAR in the anechoic room using the two different headphone types. The measured stimuli were afterwards adjusted to the results of the 2-AFC of equal loudness. The difference between the headphone level and the loudspeaker level at the eardrum are analyzed, where positive values stand for a higher eardrum pressure needed during headphone reproduction to perceive the same loudness. A three-way repeated-measures analysis of variance (ANOVA) was performed, predicting the mismatch values by the factors 'room' (at the levels 'anechoic', 'office' and 'reverberant), 'stimulus' (at the levels 'tbn250', 'tbn1000', 'tbn4000' and 'uen17') and 'headphone type' (at the levels 'HD650' and 'DT770'). Non-significant results of Shapiro-Wilk tests ($p > .05$) and visual inspection of model residuals through quantile-quantile plots suggested that the assumption of normal distribution was fulfilled. Significant main effects of Room [$F(2,12) = 26.24$, $p < .001$] and Stimulus [$F(3,18) = 14.7$, $p < .001$] were observed, while the latter must be interpreted in presence of a significant interaction between Stimulus and HP [$F(3,18) = 11.02$, $p < .001$]. The results can be found in fig. 5 where the mismatch is plotted as mean values together with two times the standard error. In the upper left corner of fig. 5 we see the influence of the room on the mismatch of headphone and loudspeaker reproduction. Mismatches in the office room are significantly different from those of the anechoic room ($p = .006$) and the reverberation chamber ($p = .001$).

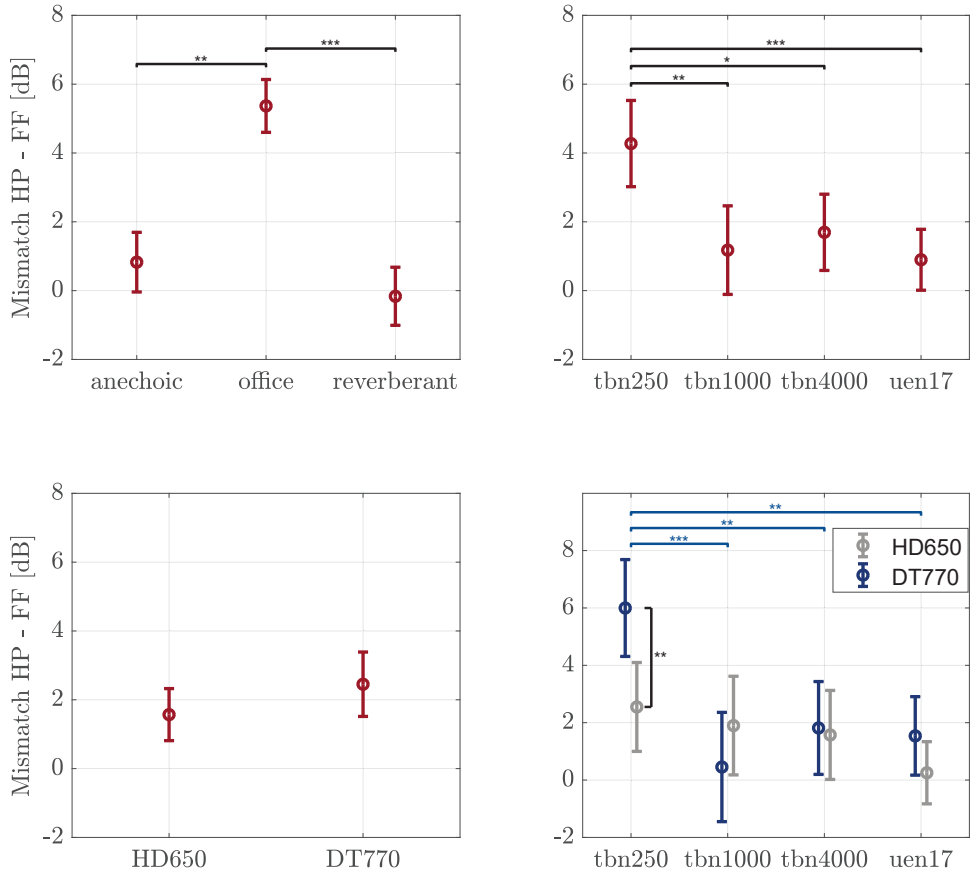

Figure 5. Results of the three way repeated measures ANVOA. Plotted are the mean values and two times standard error of each variable averaged over the other two factors (in the lower right only over the different rooms). The y-axis shows the mismatch of level when both playbacks (headphone and loudspeaker) were perceived at same loudness, positive values mean that the eardrum level was higher for the headphone reproduction. Significance levels are indicate at 0.001, 0.01 and 0.05 with three, two and one stars. The lower right plot shows the results for a simple effect analysis, where the significance level in blue (those on the horizontal axis) only show differences between the stimuli for the DT770. The black (vertical) line indicates a significant difference between the headphones.

The upper right corner of fig. 5 shows the influence of the stimuli. For low frequencies (250 Hz), an increased mismatch can be seen that differs significantly from all other (tbn1000: $p = .002$, tbn4000: $p = .048$, uen17: $p < .001$). Anyhow, these findings have to be seen in an interaction with the headphone type. While the plot in the lower left corner suggest a tendency for the DT770 to be perceived less loud (and therefore needing more amplification to excite the same loudness), the difference as such is not significant but can be further divided by the stimuli used. The lower right corner of fig. 5 shows the interaction between headphone type and stimulus. A follow up simple effect analysis divided into the effects headphone type and stimulus was done to investigate the interaction between headphone type and stimulus using a univariate analysis of variance. While no significant difference between the stimuli using the HD650 was found the use of a DT770 [$F(3,86) = 8.63$, $p < .001$] does. A tukey post-hoc-test was done to find that only the 250 Hz stimulus is significantly different

from all other stimuli using the DT770 (tbn1000: $p < .001$, tbn4000: $p = .003$, uen17: $p = .002$). Between the headphones only a significant difference [$F(1,43) = 8.61$, $p = .005$] can be found for the 250 Hz stimulus.

3.5 Discussion

The results for the stimulus and headphone interaction are mainly in accordance with those found by Fastl et al. [2] in terms of higher differences for low-frequency narrow band excitation as well as higher amplification needed for a closed headphone design to perceive the same loudness compared to an open design. Yet, the results for a 1 kHz stimulus differ. This might be due to the fact that this work investigates narrow-band noise and not pure tones. Additionally, Munson and Wiener [1] also found a higher mismatch for the closed headphone design and low-frequency excitation. The results for the reverberation chamber are not as expected. Feedback from participants indicated an 'unnatural' perception of the reverberation chamber, a disturbance by noise they made (especially clicking of the pedal switch) and whether they should concentrate on the playback itself or at the decaying process or both. Additionally, the pause between headphone and loudspeaker playback was increased, as the decaying process needs time and participants should listen to the whole decay. Therefore, the reverberant room might be unsuitable for loudness balancing as it introduces an additional dimension to the term of loudness: a separate perception of playback and reverberation and an unwanted pause. It should be noted that results indicate no interaction between the factor 'room' and other factors, and the statistical results for the other factors are not influenced by taking out the data for the reverberation chamber. Nevertheless, there is a significant difference between the mismatch in the anechoic and office room, yet more data is needed to quantify and explain the difference in future studies.

3.6 Conclusion and outlook

A listening test was conducted to a) show that the missing 6 dB can be found using the setup and to b) investigate whether the room has an influence or not. To explain the results, different methods to measure the SPL at the eardrum when using headphones were investigated and especially the fitting of the headphone seems to affect low-frequency responses of closed headphones. The results indicate a good consensus with literature data where low frequency and closed headphones lead to lower perceived loudness and additionally the results suggest a significant influence of the room, although results taken in the reverberation chamber can not fully be explained. Yet, the influence of the room cannot be neglected and should be taken into account for headphone reproduction of virtual or augmented realities.

Further listening tests are in progress at two facilities in Aachen and Oldenburg. At both sites, an anechoic chamber is used and an office-like room that differs between both sites. Therefore, two rooms apart from the anechoic rooms will be included. Furthermore, a limited number of test subjects will participate at both sites in the listening test. The eardrum levels are measured using probe tube microphones to account for individual characteristics (including the individual HRTFs) and screen individual headphone fitting. Additionally, influence of different source width and coloration perception is taken into account when comparing headphone to loudspeaker reproduction. To further investigate different reproduction methods in-ear transducer will be included. And finally, a higher number of participants will increase the statistical power of the listening test.

ACKNOWLEDGEMENTS

The authors like to thank everybody who participated in the listening test voluntarily. The project was funded by the Head Genuit Foundation under the project ID P-16/4-W and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 352015383 – SFB 1330 A4.

REFERENCES

[1] Munson, W. A.; Wiener, F. M. In search of the missing 6 dB. The Journal of the Acoustical Society of America, 24(5), 1952, p. 498-501,

[2] Fastl, H.; Schmid, W.; Theile, G.; Zwicker, E. Schallpegel im Gehörgang für gleichlaute Schalle aus Kopfhörern oder Lautsprechern. Fortschritte der Akustik, DAGA, 1985

[3] Sivian, L. J.; White, S. D. On minimum audible sound fields. The Journal of the Acoustical Society of America, 4(4), 1933, p. 288-321.

[4] Brinkmann, F.; Lindau, A.; Weinzierl, S. On the authenticity of individual dynamic binaural synthesis. The Journal of the Acoustical Society of America, 142(4), 2017, p. 1784-1795.

[5] Völk, F.; Fastl, H. Locating the missing 6 dB by loudness calibration of binaural synthesis. In Audio Engineering Society Convention 131. Audio Engineering Society. 2011

[6] Denk, F.; Kohnen, M.; Llorca Bofi, J.; Vorländer, M.; Kollmeier, B. Audiology with two ears: Does binaural hearing influence the loudness mismatch between free-field and headphone presentation, European Society of Audiological Societies Congress EFAS, Lisbon, Portugal, 2019

[7] Masiero, B.; Fels, J. Perceptually robust headphone equalization for binaural reproduction. In Audio Engineering Society Convention 130, Audio Engineering Society, (2011).

[8] Ewert, S. D. AFC - A modular framework for running psychoacoustic experiments and computational perception models, in Proceedings of the International Conference on Acoustics AIA-DAGA 2013, Merano, Italy, 2013, p. 1326-1329.

[9] Levitt, H. C. C. H. Transformed up-down methods in psychoacoustics. The Journal of the Acoustical society of America, 49(2B), 1971, p. 467-477.

[10] Zwicker, E.; Fastl, H. Psychoacoustics: Facts and models (Vol. 22). Springer Science & Business Media. 2013.