Sound Quality Evaluation of Coffee Grinders in Virtual Reality

André Kruh-Elendt[1], Ahmet Berk Selvi[2], Michael Vorländer[2]

[1] *HEAD acoustics GmbH, 52134 Herzogenrath, Deutschland, Email: andre.kruh-elendt@head-acoustics.com*

[2] *Institut für Hörtechnik und Akustik, 52074 Aachen, Deutschland, Email: berk.selvi@rwth-aachen.de, mvo@akustik.rwth-aachen.de*

Introduction

Sound is a relevant aspect of most products as it influences the user's interaction, acceptance, and overall experience. As most perceptual evaluation tasks, product sound quality (SQ) assessment is a process influenced by external and internal factors. In this regard, Blauert [1] mentions *sound source*-, *situation*-, and *individual*-related factors in particular. As part of the situation-related factors he highlights the dependence of (product) sound perception and evaluation on the context, which encompasses aspects such as aesthetics, the state of activity of the listener, and the adequacy or suitability of a sound in a given environment [2, 3].

The increased availability, affordability and functionality of head-mounted displays (HMDs) used for virtual and augmented reality applications opens up the option to perform listening experiments and, more specifically, product sound evaluation studies, using extended reality (XR) technologies. Hereby, additional information about the (visual) context can be provided and interactions can be implemented, two elements influencing the aforementioned situation-related factors. SQ evaluations in fully or partially virtual scenarios have the potential of providing repeatable, transferable, and controllable listening conditions compatible with acoustic signals obtained through simulation.

These potential advantages stand in contrast to the increased design and implementation complexity required for the execution of XR-based listening experiments, which might not be justified in every evaluation scenario. Moreover, the (ecological) validity and reliability of SQ ratings given in virtual environments need to be investigated.

Method

In a first attempt to assess the validity and comparability of SQ ratings given in a virtual reality (VR) scenario a simplified experimental setup with limited interaction possibilities, easy implementable workflow, and manageable simulation complexity was conceived. While this setting does not exploit some of the aforementioned advantages of VR, it was deemed necessary to demonstrate the validity in a simplified design to limit the influence of additional, difficult to control factors such as individual (participant) behavior.

A between-groups (i.e., between-subjects) experiment design was selected with grouping based on the following three listening conditions:

(a) Real-world SQ assessment of the actual appliance in a real environment (i.e., room). This condition functions as the reference

(b) Hearing Lab SQ assessment based on artificial head recordings and a picture of the appliance taken in the real-world environment. This condition was included for comparison, as it is a common perceptual SQ evaluation method

(c) Virtual Reality SQ assessment based on auralized appliance sounds and performed in a virtual model of the real-world environment

According to the described experimental scenario a null hypothesis was formulated as:

$\boldsymbol{H}_0$ Performing SQ evaluations in the three specified listening conditions has no influence on the obtained results

This implies that SQ ratings given in the hearing lab and VR conditions are statistically indistinguishable from those given in the real-world setting.

Based on an a priori statistical power estimation, the total number of required samples (i.e., participants) to test the formulated hypothesis was determined to be 66, leading to three equal-sized groups of 22 participants.

Participant grouping

In order to minimize the influence of different participant group compositions on SQ ratings, a questionnaire was sent together with an invitation to all potential participants in order to collect information on their age group, gender, and experience regarding the use of the selected appliance type, VR and video games. This enabled the formation of comparable participant distributions among the three different groups.

Appliance selection

The type of appliance to be evaluated was selected following constraints such as availability, variety (e.g., brands, price, etc.), accessibility to a suitable listening environment and the expected simulation complexity. The selection process lead to the choice of electrical coffee grinders as a product type. Following this decision, specific devices were acquired with variations in price, brand, and grinding mechanism as depicted in Figure 1.

All of the selected devices displayed comparable temporal behaviors consisting mainly of a grinding and an idling phase, each with different sound characteristics, and a transition phase of varying duration in-between. Thus, the appliance states to be evaluated were determined to be:

Full event 30 to 35 seconds in length, and encompassing all phases (grinding, transition, and idling)

Grinding Around 15 seconds length, with more or less

(a) BAR **(b)** BOD **(c)** DEL

Figure 1: Electric coffee grinders chosen to perform the SQ evaluation experiment. The selection was performed considering aspects such as price, brand, and grinding mechanism.

constant characteristics

Idle Steady state sound of 10 to 12 seconds in length following the grinding and transition phases

Listening environment

Based on the selected appliance type (i.e., coffee grinders), a kitchen, located inside the headquarters of HEAD acoustics, was chosen as a suitable environment for the SQ assessment process. The selected room was used to perform the artificial head recordings used in the hearing lab condition and also served as the base for visual and acoustic modeling in VR.

Stimuli for Virtual Reality

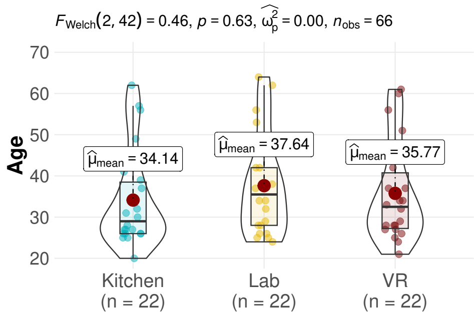
While the real-world (i.e., kitchen) and hearing lab stimuli could be reproduced without much effort, requiring a remote electrical switch for the former and calibrated artificial head recordings for the latter, the acoustic simulation for the VR experiment involved a higher effort. For this listening condition, the stimuli were generated using the Virtual Acoustics (VA) auralization framework [4] which combined free-field recordings and measured source directivities of the coffee grinders in the defined states, together with measured head-related transfer functions (HRTFs) of an artificial head, and the output from the room acoustic simulation software RAVEN [5] to render a binaural audio stream. The output was calibrated based on the artificial head recordings of the real devices and updated in real time using head-tracking information obtained from the HMD. Additionally, headphone equalization was applied to the binaural audio stream with the purpose of mitigating distortions introduced by the playback device.

Experiment procedure

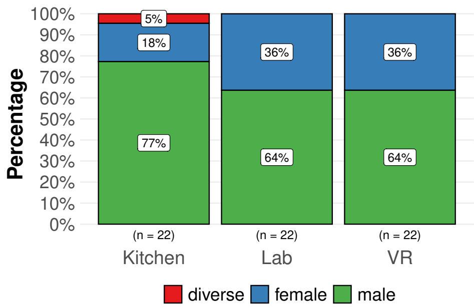
Each of the three participant groups (kitchen, hearing lab and VR) underwent a similar procedure for evaluating each of the devices (BAR, BOD and DEL) and the determined device states (Full event, Grinding and Idle). The main part consisted of a semantic differential measurement where participants rated SQ (on a 10-point scale) together with various psychoacoustic parameters (e.g., loudness, sharpness, etc.) and also provided an estimate of the perceived power and price of the appliances. The final steps of the experiment consisted of a questionnaire regarding immersion [6] and an interview to collect data about the participant. In addition, head-tracking data was recorded in the real-world and VR conditions with the purpose of observing and comparing the behavior of participants in different evaluation scenarios. However, this paper focuses on SQ and therefore only the respective ratings are reported.

Results

The final composition of participant groups showed slight deviations from the planned distributions. These are attributed to scheduling conflicts and availability during the time in which the experiments were conducted. While the age and gender distributions shown in Figure 2 are comparable, the VR participant group had a reduced number of experienced participants, as shown in Figure 3.



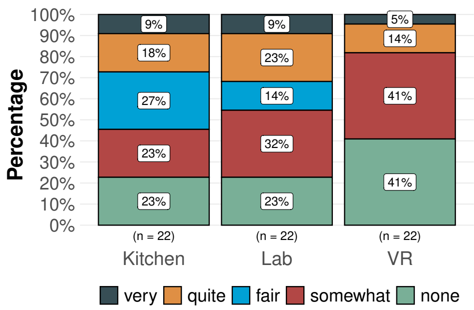
(a) Age



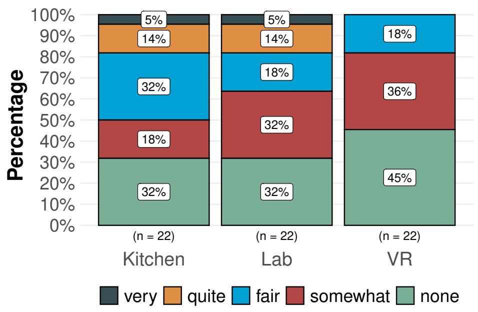
(b) Gender

Figure 2: Participant group composition. Comparable age distributions were achieved in all three participant groups. The real-world (i.e., kitchen) condition showed some minor deviations from the planned gender distributions.

Figure 4 shows the SQ ratings for all devices and device states. The conducted analysis of variance (ANOVA) shows significant differences between the three groups, which, based on a post-hoc analysis, are mainly present



(a) Self-reported experience with the participation in listening experiments



(b) Self-reported experience with product sound evaluation

Figure 3: Self-reported participant experience with (a) listening experiments and (b) product sound assessment. The distribution shows that participants in the VR experiment were generally less experienced with both.

between the VR and both other conditions (kitchen and hearing lab).

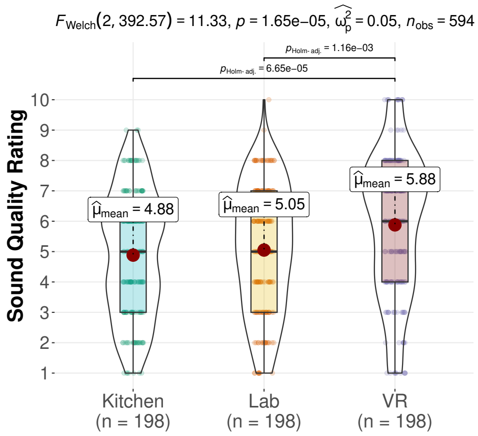


Figure 4: Sound quality ratings across the three different experiment conditions for all devices and device states. Statistically significant differences can be observed between the VR and both other experiment conditions.

A further breakdown of the SQ ratings by device and listening condition, as depicted in Figure 5, confirms these findings, showing higher overall SQ ratings in VR and, unlike the other conditions, no significant differences between any of the three evaluated devices.

Discussion

Based on these findings, further data exploration was conducted with the purpose of finding an explanation for the significantly different SQ ratings obtained from the VR experiment. First, the influence of the aforementioned disparities in experience with listening experiments and product sound evaluation between the three participant groups was inspected. In an initial approach, this was done by selecting and excluding an equal number of participants from each group until a more homogeneous distribution in the level of expertise between the subject groups was achieved. However, even after the exclusion of participants with a higher level of expertise, significant differences in SQ assessment results remained,

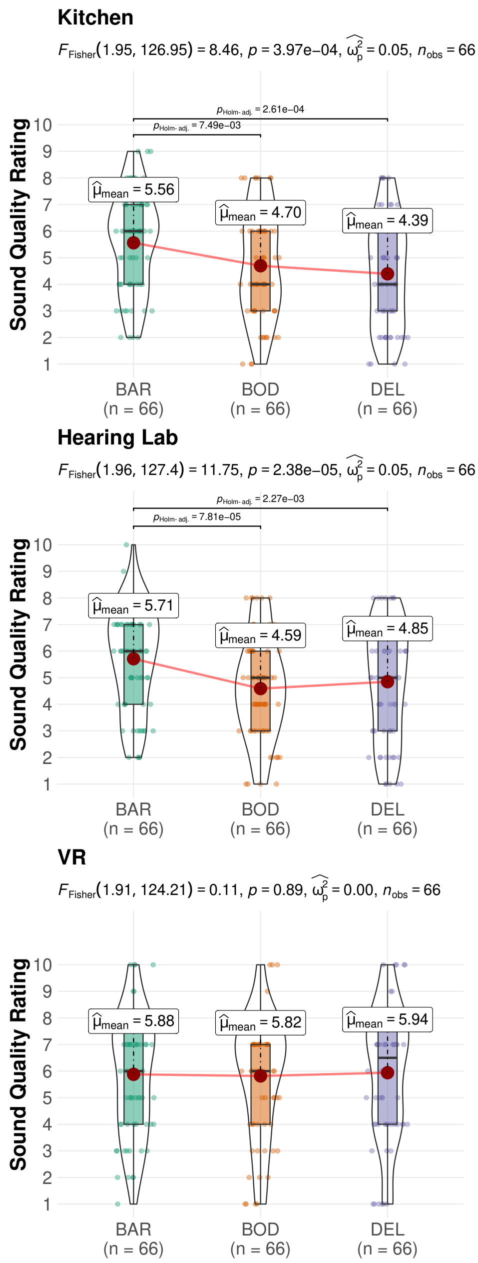


Figure 5: Sound quality ratings across devices for each of the listening conditions. In contrast to the real-world and hearing lab conditions, no significant differences between devices are observed in VR.

as shown in Figure 6.

Given the reduction in statistical power and further disadvantages arising from the selection and exclusion of individual participants, a second approach to assess the influence of participant group composition was examined. This involved training a linear regression model, determining a representative target participant profile, and comparing the impact of poststratification on SQ results. Nevertheless, this method also failed to explain the differences in SQ ratings observed from the VR condition.

A subsequent analysis of the SQ results across the three determined device states revealed that significant differences were only present for the idle device condition, as

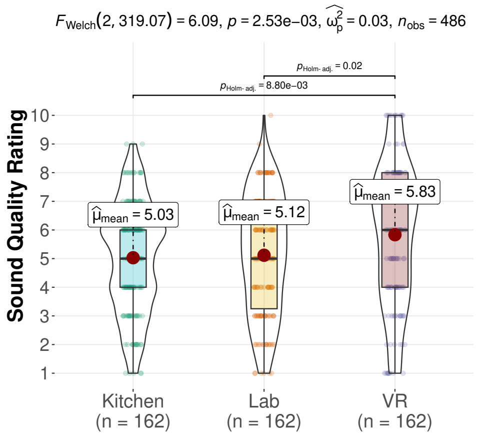


Figure 6: Sound quality ratings across the three different listening conditions for all devices and states, after balancing the participant groups based on self-reported experience with product sound evaluation. Despite the balancing, significant differences in ratings remain.

depicted in Figure 7. This observation led to a review of the auralization and playback process and the identification of the applied (idle) source directivity as the main distinct component in the reproduction chain for idle sounds in VR.

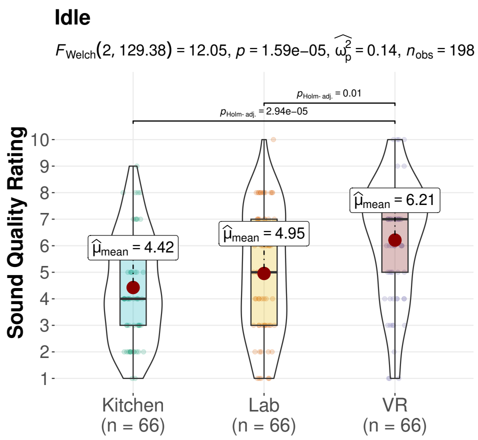


Figure 7: Sound quality ratings for the idle device state across the three different experiment conditions and all devices.

This explanation was further supported by artificial head measurements of the idle stimuli presented in VR using both the grinding and idle directivities, which exhibited noticeable differences in loudness. Furthermore, it was observed that the differences in SQ ratings between the three groups were no longer statistically significant once the idle state ratings were removed, as depicted in Figure 8.

Conclusion

The results from this study showed significant differences between the VR and both other experiment conditions leading to the rejection of the formulated null hypothesis $\boldsymbol{H}_0$. This means that the current implementation of the VR experiment failed to produce reliable and comparable SQ assessment results and therefore, additional

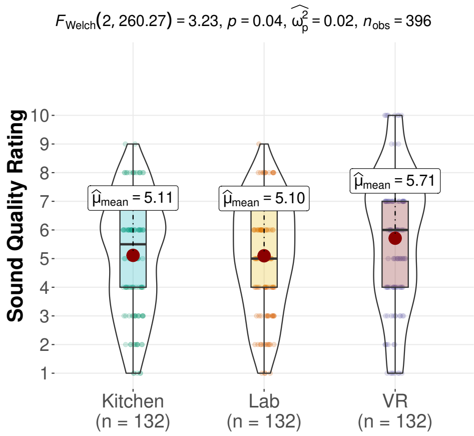


Figure 8: Sound quality ratings for all devices across the three different experiment conditions after excluding the idle device state ratings. In contrast to the results shown in Figure 4 the differences are no longer statistically significant.

optimizations are needed. In contrast to this outcome, SQ ratings obtained from the hearing lab condition were remarkably similar (i.e., statistically indistinguishable) to those from the real-world assessment, confirming the validity of product SQ ratings based on binaural recordings for the investigated scenario.
A deeper data analysis pointed towards uncertainties in source directivity measurements as a potential cause for the discrepancies in SQ ratings given in VR, which have to be confirmed and addressed in following studies, and before conducting VR experiments in more complex scenarios. This highlights the relevance of all components involved in the auralization process and the challenges faced in the implementation of realistic audio-visual experiments in VR.

Literature

[1] Jens Blauert. "Product-Sound Design and Assessment: An Enigmatic Issue from the Point of View of Engineering?" In: *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. Vol. 5. Yokohama, Japan, Sept. 29, 1994, pp. 857–862.

[2] Jens Blauert and Ute Jekosch. "Sound-Quality Evaluation - A Multi-Layered Problem". In: 83 (1997).

[3] Rainer Guski. "Psychological Methods for Evaluating Sound Quality and Assessing Acoustic Information". In: *acta acustica* 83 (1997).

[4] Institute for Hearing Technology and Acoustics, RWTH Aachen University. *Virtual Acoustics – A Real-Time Auralization Framework for Scientific Research*. URL: `http://www.virtualacoustics.org/` (visited on 03/24/2024).

[5] Dirk Schröder and Michael Vorländer. "RAVEN: A Real-Time Framework for the Auralization of Interactive Virtual Environments". In: Forum Acusticum. Aalborg, Denmark, 2011.

[6] Bob G. Witmer and Michael J. Singer. "Measuring Presence in Virtual Environments: A Presence Questionnaire". In: *Presence: Teleoperators and Virtual Environments* 7.3 (June 1998), pp. 225–240. ISSN: 1054-7460. DOI: `10.1162/105474698565686`.