Exploring the potential for Virtual reality to replace real-life sites in studies on human perception

Marcus Hamilton Siddharth Mishra Måns Nyman Johanna Simfors

Supervised by
Christopher Peters

Abstract

Previous studies have shown the usefulness of VR as a tool for evaluating many aspects of different environments. However, it is not yet clear to what extent VR can replace traditional methods of evaluation. In an attempt to shed more light on this question, we replicated a previous study concerning acoustic evaluations of different sites which compared the results of on-site evaluations and questionnaire evaluations. In our study, the on-site evaluations were replaced with VR evaluations of similar sites and a similar questionnaire was created. By comparing the results, both between the two different studies and the two different procedures used in this study alone, we explored the viability of VR in this context. We found that between the questionnaire and the VR experiment, there were no significant differences in results. Between this study and the original, the results are correlated to 65%. These results give further ground to the argument that VR could replace more costly and time-consuming methods of evaluation. Still, further studies are required in order to reach a conclusive answer on this topic.

Introduction

Virtual reality (VR) is a technology that has been commercially available since the 1980's and 90's but which, perhaps due to technological constraints, vanished from the market. Over the past decades, owing to advancements in technology and both increasing focus and demand on the gaming industry, VR has made its comeback. This experience offers an unparalleled immersion compared to e.g. a computer screen and headset. Allowing users to fully immerse themselves in different scenes and settings, it also enables stronger emotional reactions. Anyone who has tried e.g. a horror game in VR would be able to confirm that a fully immersed experience creates stronger and more realistic reactions. This creates an opportunity to explore plenty of scenarios, scripted and unscripted, that would be arduous, hazardous or nearly impossible to orchestrate in the real world.

The immersion offered by VR is not limited to the visuals surrounding the user. There is a strong correlation between auditory and visual cues. Audio affects how users perceive the visual environment and could, for instance, enhance or detract from the overall experience [1]. Similarly, the visuals also affect how the audio is perceived [2].
Although mostly known for entertainment purposes, speculations regarding the uses of VR in research and professional use have existed since its early stages [3]. This is one of the main purposes of this study, to explore whether or not VR studies on human perception could mimic and represent the real world accurately enough

to be a reliable candidate, or even a replacement.

For the reasons above, the research question explored in this paper is "*Do different ambient sounds and noises enhance or detract people's perception the same way in a natural-setting and a VR-environment?*" and is based on the study by Anderson LM, Mulligan BE, Goodman LS, Regen HZ, conducted in 1983: Effects of Sounds on Preferences for Outdoor Settings [1]

Background

The perceived effects of audio in VR environments have previously been recorded in numerous studies. Examples include audio-visual design of public spaces in urban environments [4] and the effect ambient and realistic audio has on the perception of greenspace in a VR environment [5]. Both heavily implying the strong effects audio has on the perceived environments in VR.

This study attempts to, in VR, recreate the methods used in the original study, Effects of Sounds on Preferences for Outdoor Settings [1]. The more similar the methods in this study are to the original, the more significant and relevant the comparisons of the results are. By comparing them, conclusions can be drawn regarding how a virtual reality environment compares to a real environment, when studying human perception.

The original study used two sites, a wooded setting - A botanical garden with a mixed hardwood forest. The other, a completely built site, without trees, and nearly no visible vegetation, located on a street in downtown Athens, Georgia. This information lays the foundry of what our virtual settings are based on. Important to note is that the study was conducted during the summer, which the virtual settings need to reflect.

Each setting was introduced to 10 participants. Participants were asked to complete tasks, initially a visual evaluation of the setting by a series of questions. Participants were then exposed to pre-recorded audio through headphones, and were asked to rate how the audio affected their perception of the setting, as if the sound would be regularly heard there. The answers were on a scale from 1 (most detracting) to 8 (most enhancing). The recorded auditory stimuli were selected to cover a range of sounds naturally and frequently occuring in urban, rural, and natural settings [1].

Finally, a slide-and-tape procedure was conducted for new participants. An image of the site was presented, together with pre-recorded audio. The participants were asked to rate the site on the same scale as above.

Method

Figure 1 *- Completely wooded site*

Two artificial VR-environments (sites) were created using Unreal engine version 4.25 (UE4). The first depicted a natural environment without man-made elements (Figure 1). This site was created for the purpose of this study using a forest asset pack downloaded from the UE4 Marketplace. The second site depicted a fully man-made urban environment containing no natural elements (Figure 2).

This site was manually copied from a scene in Unity version 5 which had been provided by our supervisor Christopher Peters. The experiment used two different procedures: VR-testing and questionnaire. VR-testing was done in a meeting room at KTH and the questionnaire was taken wherever the participants happened to be at the time of participation and was in the shape of a Google form. In both procedures, participants first performed a visual evaluation followed by an acoustic evaluation of the same site, after which they repeated the same process for the second site. 11 participants completed the VR-testing and 23 participants completed the questionnaire. None of the VR-testers participated in the questionnaire and vice versa.

Figure 2 *- Urban site*

The visual evaluation consisted of 11 different semantic items (opposite adjectives) e.g. “Commonplace” and “Unique”. The participants were asked to rate the site from 1-7 depending on which word they felt most accurately described the site. If they found the site to be very closely related to one of the words they would rate it 1 or 7 respectively, closely related 2 or 6, slightly related 3 or 5 and if equally related, 4. In the acoustic evaluation, the participants were asked to evaluate whether a particular sound detracted from, or enhanced the site they were currently exploring. They were asked to consider how they would react to the sound if it was to be regularly heard there. The rating was given on a scale from 1-8 where 1 meant “Most detracting” and 8 meant “Most enhancing”. All questions were identical for both procedures. For the questionnaire, instead of using VR, participants were presented with still images of the sites (one image per site). For the audio evaluation, the same images were used with added audio tracks and presented in separate 20-second embedded Youtube-videos within the questionnaire.

VR-testing began with a test leader explaining the procedure to the participant alongside an example site. This site was presented as the explanation was given and one example of both visual and acoustic evaluation was shown. Having understood the procedure, the participants were instructed to sit down at the designated testing location and put on the VR-headset, adjusting it for a comfortable fit. Once ready, participants were presented with the first site and encouraged to look around and explore it briefly, after which they were verbally asked to answer the 11 different semantic items. Upon completion of the first visual evaluation, the site was temporarily closed, in order to add the first of 10 audio tracks, and opened again shortly thereafter. The added audio track was played for 20 seconds after which the site was closed and the participants were asked to rate the sound’s effect on their perception of the site from 1-8. After rating the first sound, the same process was repeated 10 times where one of the audio tracks (birds singing) was played twice (1st and 11th) in order to examine whether the order of the stimuli affected the evaluation. Having completed the first acoustic evaluation, the participants were given the option of taking a short break or to proceed immediately with the evaluations of the second site. The second procedure was identical to the first, with the same semantic items and sound stimuli in the same order with one exception; another

sound stimuli (crickets) was played twice (2nd and 11th).

Data Analysis

The results from both procedures were statistically analyzed. The analysis was split into two parts - The first analysed the data collected in our study. The second part analyzed our data in relation to our reference study. Firstly, a two-way ANOVA test with replication was performed at 5% significance level. Sound and procedure were the independent variables, and the answer from the participant was the dependent variable. This test shows if there is a statistical difference between how enhancing one perceives the sounds in the VR procedure compared to the questionnaire. This test was followed by a one-way Repeated Measures ANOVA test, performed in three different ways. Firstly, with just sound as a repeated measurement, then with sound and procedure as a repeated measurement and lastly with sound and site as a repeated measurement. These tests help us analyse the differences in mean value and variance between these three groups. A problem with the first two repeated measures tests is that in reality, site was also a repeated measure, this problem was addressed by using a stricter significance level, 1%. This problem also occurred in our reference study, and they solved it in the same way.

Lastly, two correlation tests were performed to see how much the results from this study correlated with the reference study. The test results in a correlation coefficient between -1 and 1. A correlation coefficient of +1 indicates a perfect positive correlation, -1 a perfect negative correlation and 0 no correlation at all. The reason a correlation test was used, instead of a more powerful t-test or ANOVA test, is the lack of data from the reference study in which the data was presented as different mean values, and not as exact data points. The correlation test was made both for the procedures VR versus field study, and questionnaire versus field study.

The first part of the user test, where the user rated the visuals, was mostly to familiarize the user with the surroundings. However, just like in the reference study, a comparison of mean value and variance in how much the participants liked or disliked the sites was done. It was also investigated whether the order of sounds had an effect on the answer. This was done through a paired t-test of the two identical sounds that was repeated for every participant. For the wooden site, the sound “Songbirds” was played both as the first and 11th sound. For the urban site, the sound “Crickets” was played both as the second and 11th sound.

Results

In this section the collected data from the study is presented and analysed.

Graphs of Average Rating

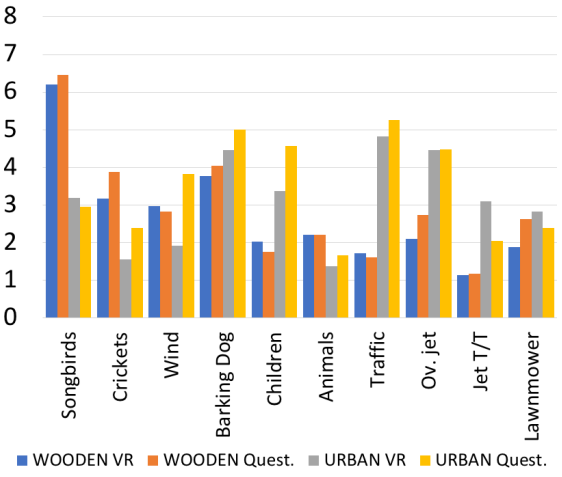

Table 1. *Average rating of enhancement in each site and procedure for each sound, where 1 equals most detracting and 8 equals most enhancing.*

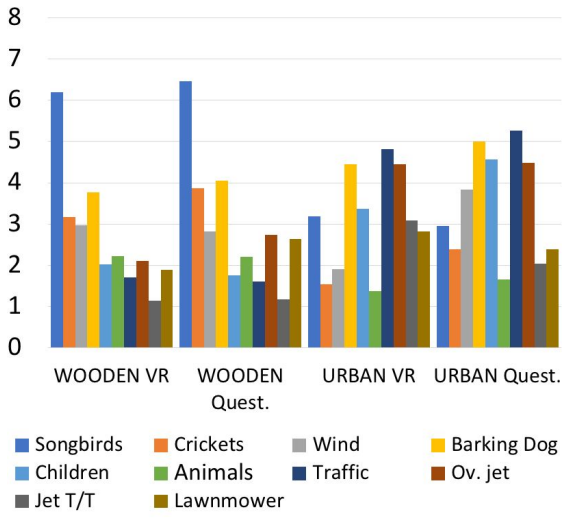

Table 2. *Average rating of enhancement in each site and procedure for each sound, where 1 equals most detracting and 8 equals most enhancing.*

A paired t-test was performed to see if there was a difference in answer between the sounds that were played twice in the same site. The test showed that there was no significant difference, indicating that the order of the sounds did not seriously distort the data.

The result indicated a strong preference of the wooden site, where the mean value was 6 both in the VR and the questionnaire. The urban site had a mean value of 3.82 in the VR and 2.96 in the questionnaire. The variance in the questionnaire was 1.82 for the wooden site and 1.95 for the urban site. The Urban VR setting had the highest variance, 3.16, and the wooden VR setting had the lowest variance, 0.8, indicating that the strongest consensus between participants was in the wooden VR setting

Statistical Analysis of Our Data
Two way ANOVA with Replication
This test analysed if there was a statistical difference between the two procedures VR and Questionnaire. The null hypothesis was that there is no difference, and the alternative hypothesis that there is. The test was conducted at a 0.05 significance level and resulted in a p-value of 0.144, a F-value of 2.146 and a critical F-value of 3.865. Since the p-value is larger than 0.05, and since the critical F-value is larger than the F value, the null hypothesis can not be rejected. In other words, it is not statistically significant to say that there is a difference between how enhancing the sounds were perceived in the VR procedure versus in the questionnaire.

One way Repeated Measures ANOVA
Three repeated measures tests were performed, with different variables as a repeated measure. The test with only sound as a repeated measures resulted in a p-value of 0.0023 (Table 3), the test with sound and procedure as repeated measures resulted in a p-value of 8.92 x 10^-9 (Table 4) and the test with sound and site as a repeated measure resulted in a p-value of 0.0022 (Table 5).

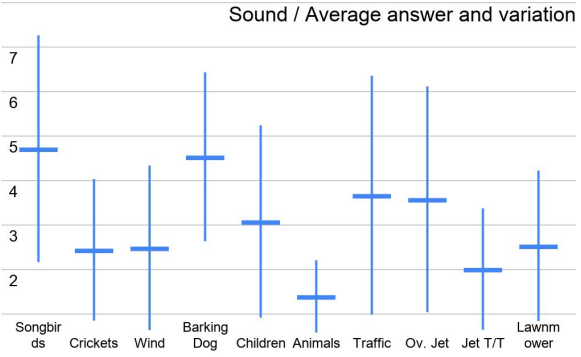

Table 3: *Average answer and variance in answer for each audio across both of the sites (wooden or urban) and both procedures (VR or questionnaire)*

The significantly lower p-value in the second test indicates that there was a bigger variance among the replies to one sound when also procedure was a dependent variable. In other words; The site used (wooden or urban) had a bigger impact on the answer than the procedure used (VR or questionnaire).

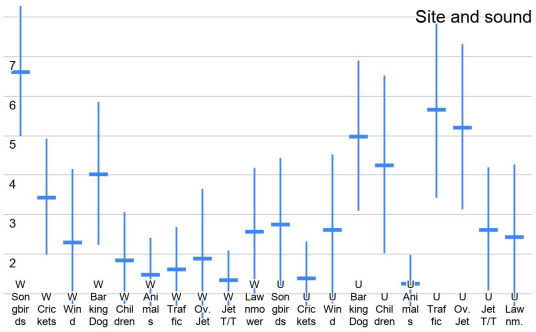

Table 4: *Average answer and variance in answer for each site and audio across both procedures (VR or questionnaire)*

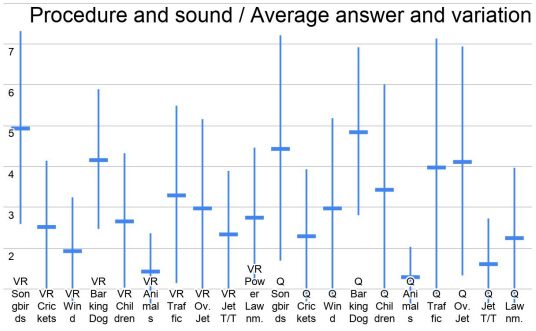

Table 5: *Average answer and variance in answer for each procedure and audio across both sites (wooden or urban).*

The sound “Barnyard Animals” had the lowest variance, which means that there was a big consensus between participants in how enhancing this sound was, no matter what site or procedure was used. The sound “Downtown traffic” had the highest variance across procedure and site. There was a big consensus that the sound was detracting in the wooden site, but the rating in the urban site differed a lot.

Statistical Analysis in Relation to Reference Data

Two correlation tests were performed to see how much the results from this study correlated with the reference study. The first measured the correlation between the mean value of the enhancement-rating for all sounds on both sites in VR versus the mean value of the enhancement-rating for all sounds on both sites in the reference field study. The result: a correlation coefficient of 0.6466, meaning the data from the field procedure and the VR procedure are correlated to 65%.

The second test measured the correlation between the mean value of the enhancement-rating for all sounds on both sites in our questionnaire versus the mean value of the enhancement-rating for all sounds on both sites in the reference field study. The test resulted in a correlation coefficient of 0.6210, meaning that the data from the field procedure and questionnaire procedure are correlated to 62%.

Discussion

Sources of error

Despite our ambition and corresponding efforts to imitate the study conducted in our main reference, a few differences between the two studies exist. Alongside other potential sources of error, these differences will, for the sake of transparency, reproducibility and to support the discussion of our results, be discussed below.

The reason why these different sources of error appeared vary somewhat but are mainly due to us either initially overlooking them and failing to recognize them in time, or consciously compromising on some aspects of the study in favor of others. The time-constraint put on this project is not meant to serve as an excuse for these sources of error but rather as an explanation to why some parts of the study were not perfectly executed.

The VR-procedure was flawed in mainly three different ways. Firstly, it was aimed at imitating the field (on-site) procedure that our main reference conducted. The biggest difference between the two procedures was that in our referenced paper, 10 different participants evaluated two different sites whereas our study had all participants

evaluate both sites. This needs to be taken into consideration when comparing the two studies and before drawing conclusions.

Secondly, three different group members took turns leading the experiment and explaining the experiment to the different participants. Even though all members structured this explanation around the one given in text format for the questionnaire, some differences in the way the explanations were given undoubtedly presented themselves over the course of the experiment. Small though they may be, these differences could make the difference in the participants' ratings. A sounder approach could have been to not give an oral explanation, but rather simply having participants read the explanation that the questionnaire-takers were presented with and answer any potential remaining questions afterwards.

The third thing initially overlooked in the VR-procedure concerns the optional breaks in between sites. These breaks varied from 0-5 minutes, depending solely on the participant. Measuring how, if at all, this affected the evaluations is not an easy thing to do (especially post-experiment). No formal interviews were conducted after the experiment, but some participants gave different types of feedback in an informal manner. Two participants explicitly stated they would not want to repeat the evaluations for a third site as they felt fatigued from the first two. Taking this into account, it is likely that several other participants felt the same way but did not mention it afterwards. Also likely is that several participants did not want to take up our time (or their own) more than necessary and therefore chose not to take as long a break as they might have needed in order to stay focused. A mandatory 5-minute (or other set time) break might have been a better approach to this.

The questionnaire procedure was aimed at imitating the slide-and-tape procedure conducted in our main reference but fell slightly short in mainly two ways. In our questionnaire, we had no way of controlling the volume when participants performed the acoustic evaluation. If the volume was either too low or too high it could highly influence their ratings. In contrast, the slide-and-tape procedure was done in a classroom with all participants hearing the stimuli at roughly the same volume (depending on where in the classroom they were sitting relative to the speaker).

Moreover, the images in the slide-and-tape procedure were shown on a projector screen while the images in our questionnaire were smaller in terms of both actual and relative size. The questionnaire-takers were not asked what type of screen they were using when taking the questionnaire, which could potentially be any type of screen and size ranging from smartphone to projector screen. This factor could lead to big differences in terms of immersiveness and affect the ratings.

On a general note, concerning both procedures in this study, we want to address the similarities and differences between the sites created for this experiment and the sites used in our main reference. For several reasons, we did not strive to create replicas of the referenced sites. Due to poor image quality in the referenced paper, complete replicas were impossible. In some regards, it would have been possible to create more similar sites than we ended up doing but we chose not to. The reasoning behind this decision is that the most important aspect of the sites was not for them to look exactly like the ones in our reference. Rather, the aspect deemed most important was that they differed greatly from one another and that they successfully portrayed a completely

urban and wooded (natural) site respectively.

Despite the slight differences between the two studies, the results indicate that conducting evaluations of this kind is feasible in a VR-setting. Just as in our referenced study, no major differences were found between the procedures. We assumed a VR-environment would be more immersive and closer to reality than a still image paired with audio tracks. As such, the evaluation of a VR-environment should give the evaluator a stronger sense of realism than by simply looking at images and hearing sounds. However, the results trigger other questions that remain to be answered. Given the fact that both our referenced study and our own exhibited only small differences in the evaluations depending on procedure, one might ask if it would ever be needed to use on-site or VR-evaluations as a procedure for this type of study. The small differences between procedures could indicate that human imagination is a powerful enough tool to simulate an environment in this manner. No such conclusions can be drawn from the results of this study alone, however.

To build on this study, we recommend evaluating differences between on-site and VR-testing to see how they differ. In our study, we have merely exchanged on-site for VR and nothing can be said about how justified this substitution is until the both procedures are tested side-by-side.

Another aspect to consider is that the original study was conducted four decades ago, where many changes in society at large and technology has happened. The views and perception might be affected by other environmental aspects, such as how we view technology, how the increased traffic has affected how we perceive the sounds, and similarly for other sounds.

Conclusions

The aim of this study was primarily to explore whether or not acoustic evaluations were feasible to conduct using VR. To answer the research question, we need to consider two things: How do the results from our study differ between VR and questionnaire, and how does this compare to the original study? Our results indicate no significant difference in the answers in the questionnaire to the answers from the VR experiment. Solely based on this, the statement that VR is an applicable method, holds. However, as the mean values match with 65% similarity to the original study, the sources of error need to be taken into consideration. A major being the lack of actual data points in the original study.

Thus, we conclude, based on our results, in measuring how ambient sounds affect the perception of a setting—natural or virtual—VR is possibly an interchangeable method to real-life field tests, but needs further testing to be certain.

References

1. Anderson LM, Mulligan BE, Goodman LS, Regen HZ. Effects of Sounds on Preferences for Outdoor Settings. Environment and Behavior. 1983;15(5):539-566. doi:10.1177/0013916583155001

2. Stéphanie Viollona, Catherine Lavandiera, Carolyn Drake. Influence of visual setting on sound ratings in an urban environment. 2002. doi.org/10.1016/S0003-682X(01)00053-6

3. Satava RM. Virtual reality surgical simulator. The first steps. Surg Endosc. 1993 May-Jun;7(3):203-5. doi: 10.1007/BF00594110. PMID: 8503081.

4. G.M. Echevarria Sanchez, T. Van Renterghem, K. Sun, B. De Coensel, D. Botteldooren
Using Virtual Reality for assessing the role of noise in the audio-visual design of an urban public space

Landsc. Urban Plann., 167 (2017), pp. 98-107, 10.1016/j.landurbplan.2017.05.018

5. M. Lindquist, B. Maxim, J. Proctor, F. Dolins. The effect of audio fidelity and virtual reality on the perception of virtual greenspace. Landscape and Urban Planning Vol. 202 (2020). doi.org/10.1016/j.landurbplan.2020.103884