<table>
<thead>
<tr>
<th>Title</th>
<th>Validation of binaural recordings with head tracking for use in soundscape evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Ooi, Kenneth; Hong, Joo-Young; Lam, Bhan; Ong, Zhen Ting; Gan, Woon-Seng</td>
</tr>
<tr>
<td>Citation</td>
<td>Ooi, K., Hong, J.-Y., Lam, B., Ong, Z. T., &amp; Gan, W.-S. (2017, August). Validation of binaural recordings with head tracking for use in soundscape evaluation. Paper presented at 46th International Congress and Exposition on Noise Control Engineering (INTER-NOISE 2017), Hong Kong, China.</td>
</tr>
<tr>
<td>Date</td>
<td>2017</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10220/44567">http://hdl.handle.net/10220/44567</a></td>
</tr>
</tbody>
</table>

© 2017 The Author(s) (46th International Congress and Exposition on Noise Control Engineering (INTER-NOISE 2017)). This paper was published in 46th International Congress and Exposition on Noise Control Engineering (INTER-NOISE 2017) and is made available as an electronic reprint (preprint) with permission of The Author(s) (46th International Congress and Exposition on Noise Control Engineering (INTER-NOISE 2017)). One print or electronic copy may be made for personal use only. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper is prohibited and is subject to penalties under law.
Validation of binaural recordings with head tracking for use in soundscape evaluation

Kenneth OOI\textsuperscript{1}; Joo-Young HONG\textsuperscript{2}; Bhan LAM\textsuperscript{3}; Zhen Ting ONG\textsuperscript{4}; Woon-Seng GAN\textsuperscript{5}

\textsuperscript{1,2,3,4,5} Digital Signal Processing Lab, School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore

ABSTRACT

Given that sounds in real-life environments and soundscapes possess direction and distance cues, it is essential to replicate these cues in virtual reality (VR) simulations used for soundscape evaluation, to ensure that the results obtained have high ecological validity. One way to replicate the desired direction and distance cues of environmental sound sources in VR simulations is to use binaural recordings coupled with head tracking for playback, instead of regular binaural or stereo recordings. The effect of using binaural recordings coupled with head tracking in VR simulations for soundscape evaluations is thus investigated. Subjects were asked to evaluate soundscapes of in-situ locations. Subsequently, they were presented with VR reproductions of the same locations in a controlled laboratory environment. The reproductions used actual audio and video captured from the in-situ location, differing only in the type of audio used for playback. The subjects were asked to evaluate the VR soundscapes experienced using the different reproduction methods in a blind test. The validation of binaural recordings with head tracking for soundscape evaluation is based on the similarity of responses as compared to the in-situ locations, as well as the quality and immersiveness experienced by the subjects.

Keywords: Soundscape, virtual reality, spatial audio, ecological validity.

1. INTRODUCTION

Virtual reality (VR) has been actively explored and validated as a method to reproduce visual and auditory environments in laboratory conditions for subjective tests. This is because the audio-visual components of VR allow for a dual modal experience that more closely mimics actual human perception (1). In a VR simulation, the user typically wears a VR head-mounted device (HMD) that tracks head movement and displays a section of an omni-directional visual stimulus corresponding to the head orientation of the user.

Previous studies have shown that congruence between auditory and visual environments is important to the perceived soundscape quality, and this applies equally to acoustic environments reproduced using VR technology (2, 3). Given that visual stimuli in VR simulations correspond to user head orientation, the accompanying acoustic stimuli must also be spatialized to achieve the desired congruence between the auditory and visual environment. However, present reproduction methods using binaural recordings have been criticized for their lack of ability to recreate acoustic environments in a holistic manner (4). This is because such recordings are unable to replicate the “spatialness”, or the full range of distance and directional cues, of a sound field (5). The information processing stage of the human auditory system is highly sensitive to interaural delays, frequency modulation, and intensity changes when localizing sound sources (6), so the problem of replicating...
accurate distance and directional cues is amplified when considering variable listener head orientations in VR simulations. Hence, the ecological validity of VR simulations is potentially reduced when non-spatialized acoustic reproduction methods are used, thus necessitating the use of spatial audio in VR simulations.

An alternative method to achieve congruence between auditory and visual environments in VR simulations is thus to make use of ambisonic recordings (7). Ambisonic recordings can be downmixed to binaural recordings using a head-related transfer function (HRTF), and these recordings can be coupled with a head tracking system that can spatialize them in real time while taking into account listener head orientation. This reproduction method is hypothesized to overcome the abovementioned weakness of using binaural recordings alone. Therefore, this preliminary investigation aims to validate this reproduction method, which uses binaural recordings coupled with head tracking in a VR reproduction of in-situ locations and soundscapes.

2. METHODS

10 participants (7 male, 3 female) were recruited for the study, all of which were tested for normal hearing (mean threshold of hearing <15dB at 250Hz, 500Hz, 1kHz, 1.5kHz, 2kHz, 4kHz, 6kHz, and 8kHz) using an Interacoustics AD629 Diagnostic Audiometer. Participants first went for a soundwalk to perform an in-situ evaluation of the soundscapes at 5 different locations. Subsequently, they experienced VR reproductions of the same 5 locations in a controlled laboratory environment, with reproductions differing only in the audio playback method. Participants then evaluated the VR soundscapes that they experienced in the laboratory relative to the in-situ soundscapes that they experienced during the soundwalk. The soundwalk procedure is detailed in Section 2.1. The audio-visual stimuli used for the laboratory experiment are described in Section 2.2, and the setup of the playback system for the various reproduction methods is described in Section 2.3. The experimental design for the laboratory session is described in Section 2.4.

2.1 Soundwalk procedure

Participants visited 5 different locations in Nanyang Technological University, Singapore, all varying in their visual and acoustic environments. Panoramic views of each of the 5 locations are shown in Figure 1. Location A is a plaza beside a construction site, location B is a canteen, location C is an open area overlooking a museum and a fountain, location D is an area beside a lake, and location E is a garden located adjacent to an expressway.

At each location, participants filled in an online questionnaire related to its soundscape. The questionnaire consists of two parts, and the responses are averaged to form a basis of comparison with the reproduction methods described in Section 2.3.

Part 1 required participants to rate the dominance of various soundscapes elements on a 5-point scale (1: Do not hear at all, 2: Hear a little, 3: Hear moderately, 4: Hear a lot, 5: Dominates completely). The soundscape elements used for the questionnaire were traffic noise, sounds from humans, water sounds, bird sounds, wind sounds, construction noise, noise from ventilation/HVAC systems, and other sounds.

Part 2 required participants to rate the acoustic environment of the location and the dominant sound source on various 7-point semantic differential scales. The semantic differential descriptors used in the questionnaire regarding the acoustic environment were: Unpleasant – Pleasant, Uneventful – Eventful, Chaotic – Calm, Monotonous – Vibrant, and Anechoic – Reverberant. The semantic differential descriptors used in the questionnaire regarding the dominant sound source were: Indistinct – Distinct, Non-directional – Directional, and Near – Far. The descriptors are chosen based on the soundscape perception model by Axelsson et al. (8), and a previous study by Kang and Zhang (9).

2.2 Audio-visual recording and stimuli

At each location, audio-visual recordings of the location were captured while the participants were completing the questionnaire described in Section 2.1. An omni-directional camera rig consisting of six action cameras (YI 4K Action Camera) arranged about a sphere, as shown in Figure 2, was used to capture an omni-directional video of the location. An ambisonic microphone (Core Sound TetraMic) attached to a recorder (Zoom F8 Multi-Track Field Recorder) was simultaneously used to capture the sound field at the location. Both the rig and microphone are placed on a tripod at a height of 1.7m from the ground, as measured from the base of the rig. Windscreens were used to minimize wind noise while the audio-visual recordings were performed.
Figure 1 – Snapshot of location (a) A: a plaza next to a construction site; (b) B: a canteen; (c) C: open area next to fountain; (d) D: open area beside a lake; and (e) E: garden next to the expressway.

Figure 2 – Omni-directional camera rig (windscreens used for recording not shown)

The recorded videos were stitched into equirectangular panoramas (Autopano Video Pro) and then post-processed (Adobe Premiere Pro CC 2017) into spherical projections for playback in a VR HMD (Oculus Rift). The stitched panoramas are shown in Figure 1. The recorded ambisonic tracks were downmixed into stereo tracks and binaural tracks. KEMAR small pinnae HRTF were used to downmix the ambisonic tracks into binaural tracks. The audio and video were synchronized by audio via synchronization cues made with a clapper at the start of each recording session, and 1 min excerpts of the combined audio-visual stimuli were used for the laboratory experiment.

A calibrated class 1 microphone (G.R.A.S. Type 40PH Free-field Array Microphone) was also
placed near the rig while recording was performed at each location, in order to measure $L_{A_{eq},1 \text{ min}}$ levels at each location. The measured $L_{A_{eq},1 \text{ min}}$ levels at locations A, B, C, D, and E were 71.5dBA, 69.1dBA, 66.5dBA, 51.7dBA, and 67.5dBA respectively.

2.3 Setup of playback system

For the laboratory experiment, all video was played through a VR HMD, and all audio was played through a pair of closed-back headphones (Beyerdynamic Custom One Pro) in a quiet room. The $L_{A_{eq},1 \text{ min}}$ levels described in Section 2.2 were used to calibrate the loudness of the stereo and binaural tracks when used for the reproduction of the soundscape at each location, through a head and torso simulator (Brüel & Kjær 4128-C-002) in a quiet room.

2.4 Design of laboratory experiment

The efficacy of three different reproduction methods for each of the five locations in the soundwalk was examined. The three different reproduction methods use the same visual stimulus for each location, but differ in the audio reproduction method. The reproduction methods are: Reproduction with stereo track, binaural track, and binaural track with head tracking (FB360 Spatial Workstation) for spatialization of audio. Hence, 15 different scenarios (3 reproduction methods $\times$ 5 locations) were created and presented to the participants in a random order, without any mention of the existence of different reproduction methods in the scenarios. After a scenario was presented to a participant, the participant would be prompted to fill in a questionnaire related to the reproduced soundscape which was experienced. Participants were free to rotate in any direction while remaining stationary, and were free to replay the audio-visual stimulus for each scenario as many times as required to complete the questionnaire.

The questionnaire consisted of three parts. Parts 1 and 2 are identical to parts 1 and 2 of the questionnaire described in Section 2.1. Part 3 required the participants to rate the reproduction on a 7-point scale in terms of the following parameters: Immersiveness (1: Not immersed at all, 7: Fully immersed); realism (1: Not realistic at all, 7: Extremely realistic); externalization (1: Inside head, 7: Outside head); overall quality (1: Very bad, 7: Very good).

3. RESULTS

3.1 Perceived soundscape elements

The mean scores for the dominance of the soundscape elements in each location for both the soundwalk and the three reproduction methods are shown in Figure 3. Correlation coefficients for the overall perceived soundscape elements between the three reproduction methods and the in-situ soundwalk are shown in Table 1. All correlation coefficients obtained in Table 1 are also statistically significant ($p < 0.05$).

The mean rating scores for the three reproduction methods correspond very closely to that of the mean rating score for the in-situ soundwalk for almost all the different types of soundscape elements in each location. The only exception is the presence of bird sounds in location B (Figure 1), which scored a mean score of 3.0 for the soundwalk but mean scores of below 1.5 for all 3 reproduction methods. This might be because location B had speakers playing loud, artificial bird sounds at intermittent timings, and the audio excerpt used for the stimuli at location 2 for the laboratory experiment did not contain those sounds.

The combined correlation coefficients between the reproduction methods and the in-situ soundwalk are also similar, thus indicating that binaural recordings coupled with head tracking can reproduce soundscape elements from the in-situ location at least as accurately as stereo recordings and binaural recordings alone. This lends support to the idea of using such a playback method to reproduce and evaluate acoustic environments in a laboratory environment.
3.2 Semantic differential analysis of soundscapes

Semantic differential charts showing the mean scores for the soundwalk and each reproduction method on each of the semantic differential scales chosen for the questionnaires are shown in Figure 4. A repeated-measures analysis of variance (ANOVA) was conducted for each of the semantic differential scales with the reproduction method as the independent variable, with the $F_{3,27}$ and $p$-values shown in Table 2. Correlation coefficients for the semantic differential scales between the three reproduction methods and the in-situ soundwalk are shown in Table 3. For the purposes of the analysis in this section, the soundwalk is taken to be a reproduction method as well, because it forms a basis of comparison with the other three VR reproduction methods.

Reproduction methods had no statistically significant effects ($p > 0.05$) on the Monotonous – Vibrant, Chaotic – Calm, Uneventful – Eventful, Unpleasant – Pleasant. This is as expected, because these descriptors do not depend on the reproduction accuracy of directional and distance cues. Reproduction methods also had no statistically significant effects ($p > 0.05$) on the Near – Far scale.
Hence, results regarding whether the three VR reproduction methods can reproduce distance cues accurately are inconclusive. It is possible that all three VR reproduction methods are capable of replicating distance cues equally accurately. However, it is more likely that this was due to the non-individualized HRTF used for the binaural recordings in the laboratory experiment, because binaural recordings have inherent limitations in replicating distance cues when non-individualized HRTFs are used (10).

In contrast, reproduction methods had statistically significant effects (p < 0.05) on the Indistinct – Distinct, Anechoic – Reverberant, and Non-directional – Directional scales. Subsequent pairwise comparison tests with Sidak correction were thus performed for the reproduction methods on these scales.

For the Indistinct – Distinct scale, there were significant differences in both stereo recordings (p < 0.001) and binaural recordings alone (p = 0.003) when compared to the soundwalk, thus indicating that both methods were unable to replicate the directional cues in the sound field accurately to allow participants to distinguish the direction of the dominant sound source. Binaural recordings with head tracking scored significantly better than both stereo recordings (p = 0.002) and binaural recordings alone (p = 0.007) on the Indistinct – Distinct scale.

For the Non-directional – Directional scale, there were only significant differences between binaural recordings alone and binaural recordings with head tracking (p = 0.009). For the Anechoic – Reverberant scale, there were only significant differences between stereo recordings and the in-situ soundwalk (p = 0.019). No significant differences (p > 0.05) were observed between binaural recordings with head tracking and the in-situ soundwalk on the Non-directional – Directional and Anechoic – Reverberant scales, lending support to the idea that binaural recordings with head tracking are indeed capable of replicating the directional cues present in the in-situ soundwalk that are required to score similar values to the soundwalk on these scales.

Lastly, binaural recordings with head tracking have the strongest correlation with the soundwalk for semantic differential scales, when compared with stereo recordings and binaural recordings alone, as shown in Table 3. The correlation is also significant at a significance level of 0.01, indicating that binaural recordings with head tracking can replicate scores obtained from an in-situ soundwalk on a variety of semantic differential scales far more accurately than the other methods. In this sense, binaural recordings with head tracking are potentially superior to stereo recordings and binaural recordings alone at providing a holistic representation of in-situ soundscapes based on semantic differentials.

Figure 4 – Semantic differential charts by location and reproduction method
Table 2 – Repeated-measures ANOVA table for semantic differential scales across reproduction methods

<table>
<thead>
<tr>
<th>Scale</th>
<th>$F_{3,27}$</th>
<th>$p =$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indistinct – Distinct</td>
<td>9.390</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>Anechoic – Reverberant</td>
<td>3.701</td>
<td>0.018</td>
</tr>
<tr>
<td>Monotonous – Vibrant</td>
<td>0.964</td>
<td>0.418</td>
</tr>
<tr>
<td>Chaotic – Calm</td>
<td>0.291</td>
<td>0.832</td>
</tr>
<tr>
<td>Uneventful – Eventful</td>
<td>0.852</td>
<td>0.473</td>
</tr>
<tr>
<td>Unpleasant – Pleasant</td>
<td>0.887</td>
<td>0.455</td>
</tr>
<tr>
<td>Non-directional – Directional</td>
<td>3.980</td>
<td>0.013</td>
</tr>
<tr>
<td>Near – Far</td>
<td>2.261</td>
<td>0.094</td>
</tr>
</tbody>
</table>

Table 3 – Correlation coefficients for semantic differential scales between reproduction methods and soundwalk (*$p < 0.05$, **$p < 0.01$)

<table>
<thead>
<tr>
<th>Location</th>
<th>Stereo</th>
<th>Binaural</th>
<th>Binaural + Head tracking</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.808*</td>
<td>0.908**</td>
<td>0.922**</td>
</tr>
<tr>
<td>B</td>
<td>0.345</td>
<td>0.169</td>
<td>0.455</td>
</tr>
<tr>
<td>C</td>
<td>0.993**</td>
<td>0.928**</td>
<td>0.957**</td>
</tr>
<tr>
<td>D</td>
<td>0.539</td>
<td>0.751*</td>
<td>0.749*</td>
</tr>
<tr>
<td>E</td>
<td>0.440</td>
<td>0.798*</td>
<td>0.922**</td>
</tr>
<tr>
<td>Combined</td>
<td>0.679**</td>
<td>0.726**</td>
<td>0.809**</td>
</tr>
</tbody>
</table>

3.3 Quality of reproduction methods

The mean scores for the immersiveness, realism, externalization, and overall quality of the three reproduction methods by location are shown in Figure 5. The method using binaural recordings coupled with head tracking obtained the highest mean scores when averaged over all locations for all four parameters.

Furthermore, a repeated-measures ANOVA was conducted for each of the four parameters with the reproduction method as the independent variable. Reproduction methods had statistically significant effects on immersiveness ($F_{2,18} = 14.355$, $p < 0.001$), realism ($F_{2,18} = 7.324$, $p = 0.002$), and overall quality ($F_{2,18} = 8.453$, $p = 0.001$). However, the effect of reproduction methods on externalization was not statistically significant ($F_{2,18} = 1.410$, $p = 0.254$). This could be because externalization effects are sensitive to the HRTF of each individual, and the HRTF used for downmixing the audio tracks was not individualized for each participant (11).

Subsequent pairwise comparison tests with Sidak correction revealed that binaural recordings with head tracking scored significantly higher for immersiveness than both binaural recordings alone ($p = 0.018$) and stereo recordings ($p < 0.001$). Binaural recordings with head tracking also scored significantly higher for overall quality than both binaural recordings alone ($p = 0.067$) and stereo recordings ($p = 0.001$). Lastly, binaural recordings with head tracking scored significantly higher for realism than stereo recordings ($p = 0.001$), but not for binaural recordings alone ($p = 0.542$). The overall results therefore indicate that binaural recordings with head tracking can achieve the desired congruence between the auditory and visual environment that is necessary for good-quality acoustic reproduction.
4. CONCLUSIONS

Binaural recordings with head tracking have yet to be thoroughly examined as a method for soundscape reproduction and evaluation. Hence, the capabilities of binaural recordings coupled with head tracking were investigated based on their accuracy in replicating soundscape elements, similarities in ratings on semantic differential scales, and their overall quality and realism, as compared to actual in-situ locations. Binaural recordings, when combined with head tracking, were found to replicate soundscape elements accurately across locations with varying acoustic properties, and were found to be superior to both binaural recordings alone and stereo tracks in delivering a high-quality, immersive environment for the listener. However, the use of non-individualized HRTFs when downmixing ambisonic recordings to binaural recordings revealed limitations regarding the reproduction method with respect to the replication of distance cues and externalization of sound. Therefore, while the findings of this preliminary study of binaural recordings with head tracking support their ecological validity when used in soundscape evaluation, the study can be expanded with more participants, and further investigation needs to be performed with individualized HRTFs for binaural recordings.

REFERENCES

4. Aletta F, Kang J, Axelsson Ö. Soundscape descriptors and a conceptual framework for developing