PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Evaluation and comparison of novel music experiences in augmented reality

Arto LEHTINIEMI[1]; Jussi LEPPÄNEN[1]; Henri TOUKOMAA[1]; Antti ERONEN[1]

[1] Nokia Technologies, Finland

ABSTRACT

This paper evaluates and compares different ways of presenting and experiencing music in six degrees of freedom (6DoF) augmented reality (AR). A musical work specifically composed and arranged for 6DoF AR consumption was presented to test subjects in three ways; as a stereo downmix from virtual loudspeakers, as audio objects and as interactive (movable) audio objects. The audio objects and virtual loudspeakers were visualized as icons in the AR scene to help the user to move and interact with the objects. A qualitative and quantitative evaluation was conducted with 12 participants to evaluate the quality of each presentation and experiencing technique. The participants found the experiences with audio objects to be innovative and most captivating. There was clearly a bigger effort and higher mental demand in accomplishing tasks with the interactive audio objects when compared to the virtual loudspeakers. Both of the audio object experiences were found to be more natural compared to the virtual loudspeaker experience. The overall results indicate that generation of rich and specifically tailored content for new realities is valuable and was perceived well by the test subjects.

Keywords: Music, 6DoF, Augmented Reality

1. INTRODUCTION

The rise of new realities such as AR and Virtual Reality (VR) offer new possibilities for rich media experiences. Streaming music services have already changed the way we consume music. Using such services, music is quickly accessible regardless of time, place or personal storage limitations. Volumetric or six-degrees-of-freedom (6DoF) AR/VR brings a new level of immersion to music consumption. So far there has been very little volumetric musical content available, especially content that is solely composed and arranged to be used in a 6DoF environment. This is due to the lack of standards for representing and rendering such content. This paper studies the experiencing of a volumetric musical work in three different ways in augmented reality. The three AR experiences were evaluated in a user study with 12 test subjects.

2. RELATED WORK

Augmented reality audio experiencing can be done with open back headphones allowing acoustic transparency for the real world or with loudspeakers on the headband of a head-mounted device not blocking the ears such as in the Microsoft HoloLens devices (1). There are also prototype head-worn AR devices which involve capture of the real audio environment and rendering it with minimal latency to the user, augmented with synthetic content (2). The benefit of such devices is that they also allow control of the real-world sounds, however, the processing needs to happen with extremely low latency.

Lokki presents some applications for augmented reality audio (3). Examples include notifications linked to places in the physical environment; application information such as calendar notifications linked to the location of the user, not a physical location in the environment; modifying environmental acoustic noise to emphasize certain events; and auditory telepresence.

Despite the growing commercial interest towards mobile AR, few if any studies have systematically investigated these audio experiences. Especially, music experiencing in AR is little studied. In a non-mobile scenario with physical loudspeakers, Mueller et al. have studied mid-air

1 firstname.lastname at nokia.com

interaction with audio objects (4). For audio rendering and user position tracking, they used Wave Field Synthesis using a large array of loudspeakers and computer vision tracking, respectively. Their system enables direct interaction with sound objects as an alternative to indirect interaction mediated with controllers or visual interfaces. The authors propose that the system can be used as a spatial music mixing room.

In a mobile AR scenario, Yang et al. studied synthesizing spatial sounds from arbitrary real objects and rendering them directly to the user's ear pods (5). Their study evaluated the usability and usefulness of the system by testing the localization accuracy of users. The authors also suggest that the sense of immersion for AR reproduction of notification sounds can be improved by including an environment model which adjusts the rendered reverberation time based on an ideal model of the environment.

Some studies present methods of incorporating rendering time acoustics to AR audio rendering. Audfray et al. propose an acoustic scene programming model for augmented and mixed reality (6). In their model, source properties such as position trajectories and directivities are preauthored while the reverberation properties of the rendering room are specified during rendering time.

Jot and Lee propose an adaptation procedure for modifying a reference binaural room impulse response (BRIR) to match the qualities of the reproduction room (7). The reference BRIR is truncated and the diffuse reverberation tail is modeled with an artificial reverberator with parameters adjusted so that the reverberation characteristics model the rendering environment. The level of the early part of the BRIR is scaled according to the local room volume.

Erkut et al. update the reflection paths of a scattering delay network to match the rendering environment geometry (8). Their ongoing work aims to extend their system for rendering acoustics for outdoor scenes.

3. SYSTEM DESCRIPTION

Our system was implemented as a distributed system running on two devices: a personal computer and a mobile phone. The personal computer hosted the audio signals (objects or channels), maintained the virtual scene state, and provided the content stream to the mobile device. Custom software was implemented for maintaining the virtual scene state such as object position and orientation metadata, linking those to audio signals streamed from a digital audio workstation (DAW), and providing a stream of audio content and metadata to the client (Figure 1).

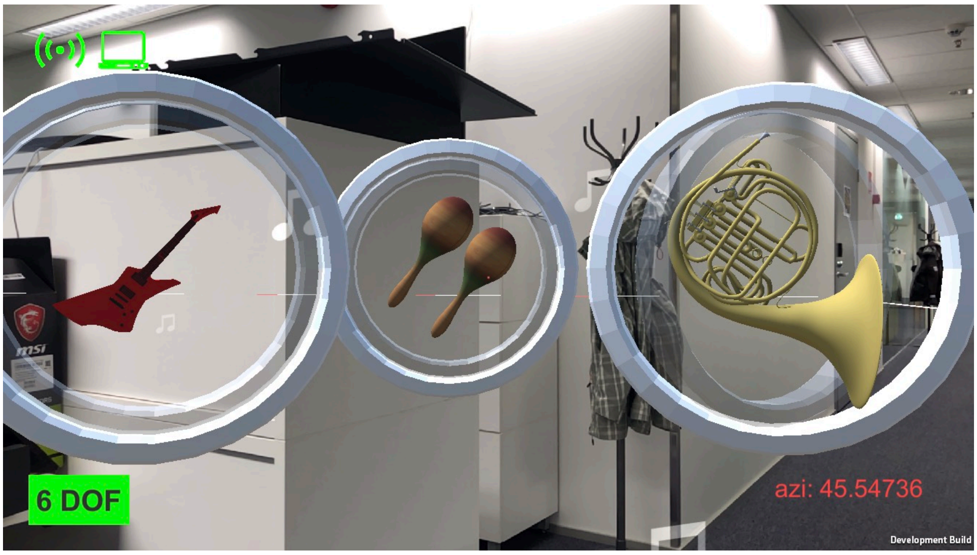

Figure 1 – The iPhone client visualizing instrument positions as icons

The mobile phone was an Apple iPhone which enables tracking the user position and orientation with the ARKit (9). The mobile phone was running head-tracked binaural rendering to reproduce spatial audio to the users via headphones connected to its headphone jack. Binauralization was done by convolving with non-personalized head-related transfer functions (HRTF). The distance rendering for objects utilized dry/wet mixing between a dry object signal and a captured wet signal in addition to distance attenuation modeling for creating a distance percept. In this study, we did not adapt the reverberation characteristics to the real environment. Analysis of the importance of modeling real world environment acoustics for the different presentation techniques is left for future work. The mobile client rendered the visuals of the scene instruments as icons to the user overlaid with a video feed of the environment, and enabled interacting with the objects via the touch UI (Figure 1).

4. VOLUMETRIC MUSICAL WORK

A pop/rock song called 'Top' was specifically composed and arranged for AR/VR consumption. The problem in using traditional multi-track content is that it is originally produced for a single listening point (i.e., a stereo downmix) where the balances between the different instruments are fixed. While it is possible to place these tracks into a volumetric scene, moving inside the scene often makes the listening experience suboptimal. In this case the balances between the instruments change and the overall mix might not sound good. In order to create immersive and volumetric experiences, the user should be able to move inside the sound scene to experience it from different locations. However, without careful design of the volumetric musical work, the audio experience may not be optimal at all locations.

In order to overcome this problem, the main design principles for the volumetric musical work creation were:

- User movement is encouraged by providing additional content at different regions of the scene
- Additional content should encourage the user to experience the song multiple times
- There should be multiple locations in the scene where the mix sounds reasonably good (according to the artist preference)
- Clear contrast between the multiple locations or "zones" in terms of playing style and sounds
- Cross-talk between different content inside the 6DoF scene should sound good, i.e., the neighboring content items should sound good when played together
- There should be clear baseline instruments keeping the song together and audible everywhere

This approach resulted in creating a "multiple sweet spot" content, where baseline instruments are in the middle and in addition there are multiple regions or "zones" with different instruments. The user can explore the different regions to enhance the experience as the harmonic arrangement changes due to 6DoF rendering. The possibility of using sound source directionality was taken into account to enhance the effect of the different zones.

The composed song was orchestrated and recorded into a volumetric experience as presented in Figure 2. During the recording, separately recorded instrument tracks were played back through several loudspeakers placed in a room. Each instrument track was played from a single loudspeaker. The audio for the volumetric experience was then recorded using close-up mics placed next to each loudspeaker. This setup mimicked a live recording setup, where each loudspeaker represented an instrument.

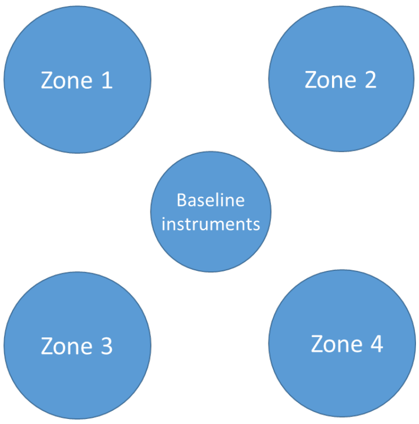

Figure 2 – Illustration of the composed "multiple sweet spot" musical work arrangement (left) and the actual recording arrangement (right)

In addition to the volumetric experience described above, a stereo downmix was created. The artist selected most of the instrument tracks and created a good sounding overall downmix of the song that incorporates elements from all of the zones.

5. RESEARCH APPROACH AND METHODS

The research topic was approached through three main research questions:

RQ1: How do the test subjects perceive the three different ways of consuming a musical work?

RQ2: Which presentation of the content performs best with the test subjects?

RQ3: How did the test subjects consume the volumetric musical work?

To answer these questions, a qualitative and quantitative user study with 12 participants was arranged. Each experiment consisted of three sessions of three different ways of presenting and experiencing music in 6DoF AR. The following three 6DoF AR scenes were used in the sessions:

1. Stereo: a scene with two virtual loudspeakers rendering the stereo mix of the volumetric musical work.
2. Audio Objects: a scene where the instruments were represented with audio objects at different positions in the scene.
3. Interactive Audio Objects: same as the 'Audio Objects' scene, but having a possibility of repositioning the audio objects via user interaction.

Each test subject participated in all three sessions, during which they were tasked with finding a preferred listening point, i.e., the position and direction where the content sounded best to them. The test subjects were first instructed to walk around the whole test area before making their decision on the preferred listening position to make sure the whole test scene had been experienced before making the selection. Before each test, the test subjects were given a brief verbal description of what to expect in the scene (audio objects vs virtual loudspeakers). For the Interactive Audio Objects scene, the test subjects were given instructions on how to manipulate the audio objects during the test. No time limit was set for the task. The test was conducted in a large open space, roughly 8m by 8m in size, where there was plenty of room to move without any obstacles. Half of the test subjects experienced the Stereo scene first and the other half the Audio Objects scene.

The used research method was a combination of observation and a selection of questionnaires. The questionnaires were: Visual Symptoms Questionnaire (VSQ), Simulator Sickness Questionnaire (SSQ), AttrakDiff, NASA TLX (task load index), and a small customized questionnaire with open questions. (10, 11, 12, 13).

Each subject was asked to fill pre- and post-test VSQs and SSQs to verify the presence and severity of possible side effects before and after each session in the experiment. We decided to use VSQ and SSQ because we wanted to compare this AR experiment to VR experiments in the future.

The AttrakDiff questionnaire was used to measure the attractiveness of this interactive system and experience. NASA TLX was used to measure the participants' workload in six categories: Mental-, Physical- and Temporal-demand; Performance; Effort and Frustration. The small customized questionnaire included feedback related to the appeal and sound quality and open feedback regarding the experiences. Each questionnaire was filled before and after each session (Stereo, Audio Objects and Interactive Audio Objects).

All 12 test subjects were male audio-oriented engineers working for Nokia and were aged from 30 to 45 years old. The participants were selected using convenience sampling where the sample is drawn from that part of the population which can be reached easily and conveniently. Participation was voluntary with a reward of two movie tickets per test subject. Almost all test subjects had a music related hobby, mostly guitar or keyboards.

6. RESULTS

In general, the three different experiences were well perceived by the test subjects and they were considered to provide new media experiences (RQ1).

6.1 Preferred listening position selection

In order to understand how the test subjects consume music in 6DoF, the selection of the test subjects' preferred listening positions in each of the scenes was monitored (RQ3).

Figure 3 shows the preferred listening position found by the test subjects. In the Stereo scene rendered with two virtual loudspeakers, the preferred listening position for all test subjects lies on a line perpendicular to a line between the loudspeakers (Figure 3, left). Furthermore, in all the cases the test subjects were facing the loudspeakers in their preferred listening positions.

When the content was presented as audio objects, the preferred listening positions were more varied (Figure 3, middle). The test subjects picked positions in different parts of the scene, mostly between the baseline instruments in the center and the instrument groups on the outside of the scene. Two test subjects, 3 and 5, selected positions near the edge of the scene. For every test subject, the direction in the preferred listening position was towards the center of the scene. Viewing towards the center of the scene, where the vocals, bass and cajon were located, results in a mix that resembles a traditional stereo mix of a song, where these instruments would often be placed near the center.

Figure 3 (right) shows the preferred listening positions for the Interactive Audio Objects scene where the test subjects could move the audio objects to different positions. Note that the positions of the audio objects before they were moved are shown. Some test subjects chose a preferred listening position close to the one they selected for the non-interactive audio objects scene; five test subjects were within half a meter from the non-interactive scene preferred listening point. Other users commented that as the content was rich and there were many good listening points, they wanted to experience a slightly different listening position for this task. Selected instruments were moved to further enhance the experience. The largest difference in the preferred listening points was approximately 4 meters. Most test subjects' viewing direction was again towards the center of the scene. Generally, the test subjects moved the audio objects closer or further away from themselves to adjust the mix at the preferred listening position more to their liking. Some test subjects repositioned the audio objects so that their positions represented a traditional live band setup, i.e., all audio objects in front of the test subject with the vocals and percussion in the middle and the other instruments to the sides. The test subjects' musical hobbies tended to influence how the audio objects were moved. Some test subjects with a guitar playing background concentrated on repositioning guitars while the test subjects with experience on keyboard playing concentrated on the synths.

Each test subject went through all three sessions during the same experiment which lasted from 40 to 60 minutes. Experiencing the Interactive Audio Objects scene took the longest time in average, ~20 minutes per test subject, whereas the other sessions lasted around 11 to 13 minutes in average.

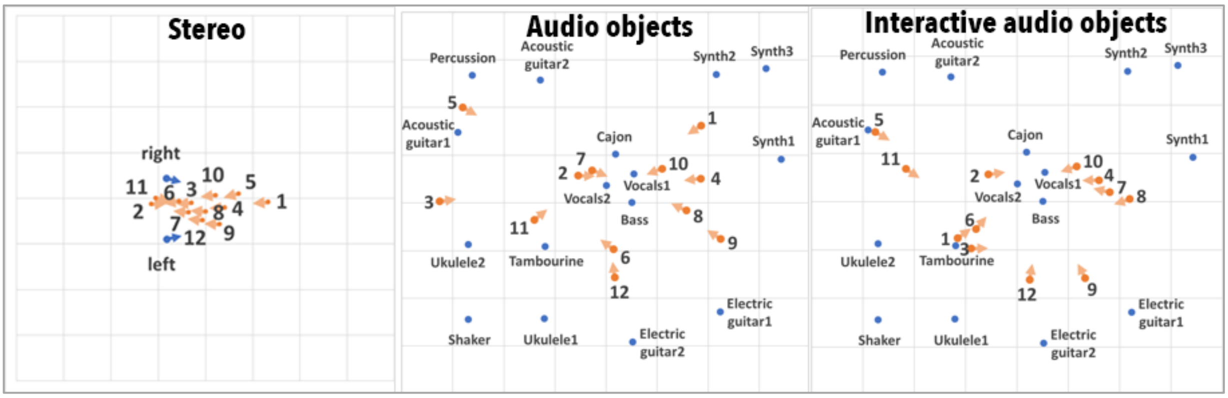

Figure 3 – Preferred listening positions for Stereo, Audio Objects and Interactive Audio Objects, respectively, providing insights into RQ3 (The grid cells are 1m by 1m in size)

6.2 Attractiveness of the system

The AttrakDiff questionnaire was conducted to learn about the attractiveness of the three different ways of consuming the musical work in AR (Figure 4). In the questionnaire the word-pairs were in random order and half of them were in reverse scale.

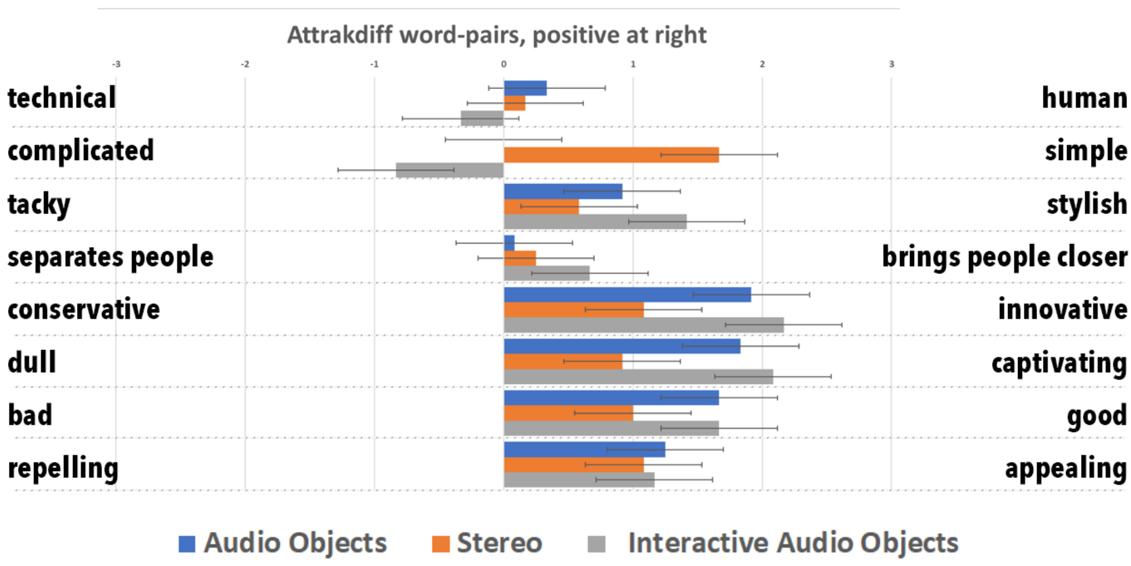

Figure 4 – AttrakDiff word pairs were rated by the test subjects (RQ1, RQ2)

Most of the statements were ranked to the positive side. The Interactive Audio Objects (i.e, the scene where the test subject was able to move the sound sources within the scene) was found slightly technical and complicated. This could be explained by the implemented interaction method for moving the objects using the mobile device touch display. The test subjects were able to perform the task but this particular action had a slightly bigger learning curve compared to experiences where the test subjects only physically moved around the scene. Despite the higher learning curve, the Interactive Audio Objects experience was rated to be the most captivating and innovative due to the possibility of fully customizing the scene (with the cost of extra effort). All of the tested experiences were found good and appealing.

6.3 Perceived workload

Results from the NASA TLX questionnaire show that there is a bigger effort and higher mental demand in accomplishing tasks with the Interactive- and Audio Objects when compared to the Stereo (Figure 5). Despite of these two results the test subjects didn't feel more frustrated, this is supported by their positive comments for Interactive Audio Objects. Also, the test subjects performed equally well with all three evaluated experiences.

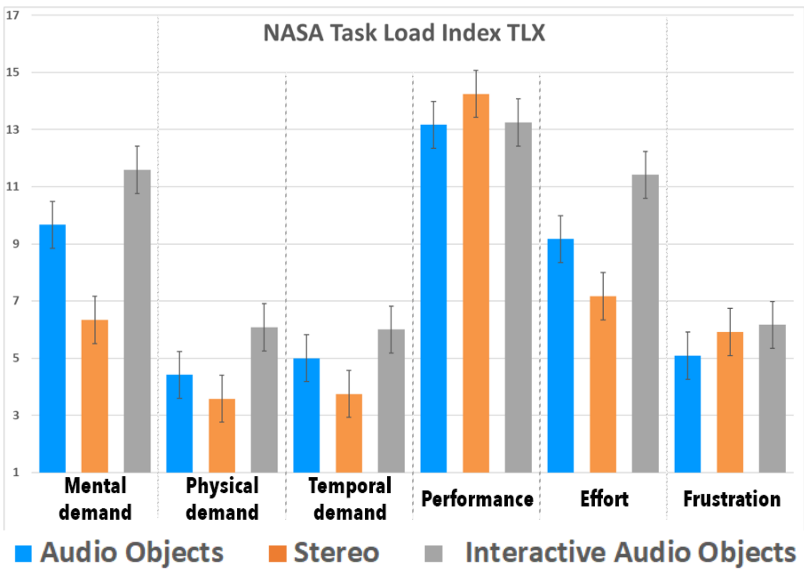

Figure 5 – NASA TLX results measuring the test subjects' workload in six categories

A bigger effort is observed when comparing Interactive Audio Objects to Audio Objects. When comparing Interactive Audio Objects to Stereo, a bigger physical and temporal demand is observed. These results were also supported by the AttrakDiff questionnaire.

6.4 General questionnaire results

One questionnaire had three questions: "I was completely captivated by the scene" with 7-point Likert scale from Strongly disagree to Strongly agree; and "Scene naturalness" and "Sound quality" both with scale attributes from poor to Excellent. Stereo results were lower in all questions (Figure 6). All answers were on the positive side and thus the negative scale is not illustrated in Figure 6.

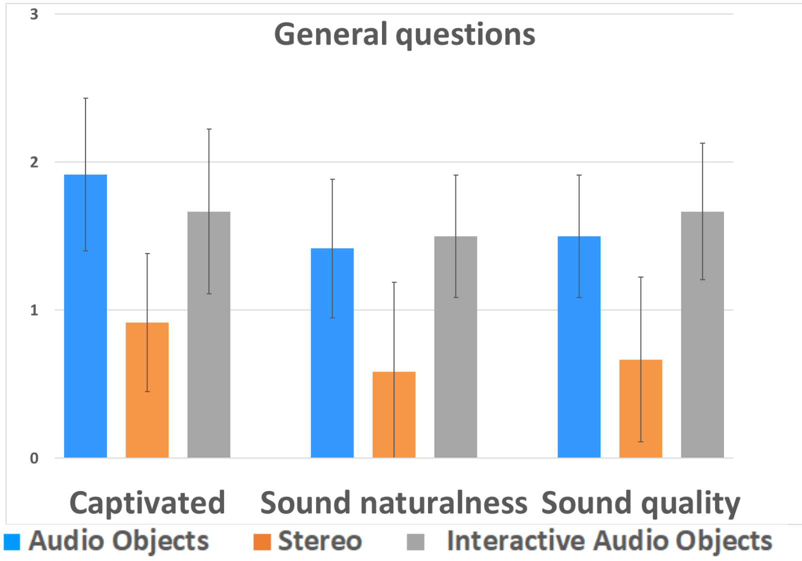

Figure 6 – General questionnaire results regarding the experiences (RQ1, RQ2)

The test subjects could also freely comment their experiences with the system. The Stereo session was said to be: "*Simple, faster to 'get in to', but not as exciting and new*". Comments related to the Audio Objects and Interactive Audio Objects were positive and describing the experience as being more interesting than traditional music listening. Most of the test subjects commented positively regarding the opportunity to move around in and explore the rich audio scene.

Interactivity was considered as a clear advantage in customizing the musical experience based on user preferences as the scene contained many alternative instruments to choose from. Moving the sound sources inside the scene was considered fun regardless of the extra effort and the Interactive Audio Objects was considered as the most interesting presentation of the musical content in this test (RQ2).

6.5 Physical considerations

Only one of the participants experienced a slight VSQ symptom, dryness. Only five participants had slight SSQ symptoms before the experiment, either a general unpleasant feeling or tiredness. One subject had both. One subject had 5 symptoms including problems in focusing and in concentrating, feel of pressure in the head and the two earlier mentioned symptoms. None of the symptoms got worse after the experiment (nor between the sessions), neither the VSQ nor SSQ symptoms.

7. CONCLUSIONS

Three ways of experiencing an original volumetric musical work specifically created for 6DoF consumption were evaluated in AR using a mobile device: Stereo, Audio Objects and Interactive Audio Objects. A user study with 12 subjects was conducted to find out how the test subjects perceive the three different ways of consuming the musical work, which presentation was considered the best and how did the test subjects consume the musical work in 6DoF in general. All of the three presentations were found to be good, but the object-based experiences were preferred over the stereo presentation due to the ability to move around in the scene amongst the instruments and explore the rich content in distinct physical locations. The Interactive Audio Objects presentation was perceived as the most innovative and captivating, and generally the best out of the three experiences. Details regarding the test subjects' preferred listening positions within each of the three presentations were described, and these can help in designing 6DoF musical experiences and systems. The design principles of creating a volumetric musical work with multiple sweet spot were outlined, including a concept of separating baseline instruments and complementing content zones in a volumetric arrangement.

Potential avenues for future work based on the user feedback and observations include utilizing the real scene acoustics into the experience and measuring the experience when the content is consumed in different acoustical environments and at different sized physical spaces. Due to the rich nature of volumetric musical works like the one designed for this study, it would be valuable to research multi-user scenarios and co-experiencing volumetric musical works. This should open new doors to even greater experiences that what is available today.

REFERENCES

1. Microsoft HoloLens 2. https://www.microsoft.com/en-IE/hololens; Online. Accessed 2019-05-23.
2. Härmä A, et al. Augmented Reality Audio for Mobile and Wearable Appliances. J Acoust Soc Am. 2004;52(6).
3. Lokki T, et al. Application Scenarios of Wearable and Mobile Augmented Reality Audio. 116th Convention of the AES; May 2004.
4. Mueller J, et al. The BoomRoom: Mid-air Direct Interaction with Virtual Sound Sources. ACM CHI Conference on Human Factors in Computing Systems; CHI 2014; April-May 2014.
5. Yang J, et al. Hearing is Believing: Synthesizing Spatial Audio from Everyday Objects to Users. Proc 10th Augmented Human International Conference; March 2019.
6. Audfray R, et al. Audio Application Programming Interfaces for Mixed Reality. 145th Convention of the AES; Oct 2018.
7. Jot J-M, Lee K-S. Augmented Reality Headphone Environment Rendering. AES Conference on Audio for Virtual and Augmented Reality; Sep-Oct 2016.
8. Erkut C, et al. Mobile AR In and Out: Towards Delay-based Modeling of Acoustic Scenes. IEEE Conference on Virtual Reality and 3D User Interfaces; 18-22 March 2018.
9. Apple ARKit. https://developer.apple.com/arkit/; Online. Accessed 2019-05-27.
10. Howarth PA, Istance HO. The association between visual discomfort and the use of visual display units. Behaviour and Information Technology. 1985;(4):135-149.
11. Kennedy R, Lane L, Berbaum K, Lilienthal M. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. Int J Aviat Psychol. 1993;3(3):203-220.
12. Hassenzahl M. The Interplay of Beauty, Goodness, and Usability in Interactive Products. Human-Computer Interaction. 2004;19(4):319-349.
13. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock PA, Meshkati N, editors. Human Mental Workload. North Holland Press; 1988. p. 139-183.