Co-Sound: An interactive medium with WebAR and spatial synchronization

Kazuma Inokuchi, Manabu Tsukada, Hiroshi Esaki

Co-Sound: An interactive medium with WebAR and spatial synchronization

Kazuma Inokuchi[1], Manabu Tsukada[1], and Hiroshi Esaki[1]

The University of Tokyo {ino, tsukada, hiroshi}@hongo.wide.ad.jp

Abstract. An Internet-based media service platform can control recording processes and manage video and audio data. Furthermore, the design and implementation of an object-based system for recording enable the flexible playback of the viewing contents. Augmented Reality (AR) is a three-dimensional video projection technology. However, there are few examples of its use as a method for audio-visual media platforms. In this study, we propose Co-Sound, which is designed as a multimodal interface that renders object-based AR dynamically in response to various actions from viewers on a web browser by sharing AR objects among multiple devices in real time. We confirmed that the system was developed as an object-based interactive medium with AR, achieved the general acceptance of the system was very high through a questionnaire survey, and low-latency synchronization to accept operations from multiple users in real time.

Keywords: Interactive media · Object-based audio · Augmented Reality · Software defined media

1 Introduction

With the spread of the high-capacity communication environments, video streaming services have expanded rapidly, and 360-degree video streaming also has attracted increasing interest. Despite the growing demand for live musical performances and concerts, it is difficult for users to view the content of package media and live broadcasting from a free viewpoint because of the limitations of the recording devices' performance and location. Few media can accept actions from viewers, as they only record and playback predesigned video and audio positional relationships as well as viewpoints.

Sound recording and playback systems can be broadly divided into three categories [3]. Object-based audio (OBA) has the following characteristics [8]: (i) Multiple objects that exist in multi-dimensional space, (ii) Interactive reproduction personalized to users, (iii) Decoupling media data from recording devices, and delivering in a variety of formats via the Internet. Unlike conventional channel-based audio and scene-based audio, an object-based approach is adopted not only in the audio, but also in other media components, such as videos and position data of instruments. The complete media data can be controlled and

managed by abstracting a series of processes from recording to playback, and OBA can interpret and express viewing objects existing in the real world.

In this paper, we present Co-Sound, an interactive audio-visual medium with WebAR. Such an audio-visual media platform is ideal for reproducing software-managed object audio. By measuring the real-time response of multiple people to the system and the QoE (Quality of Experience) of the application using this system, we confirmed that Co-Sound create new and enhanced user experiences. The main findings of these experiments were that the delay of spatial synchronization with WebRTC was lower than that with WebSocket and the accuracy of AR-marker detection and calibration could deteriorate the QoE even when the WebAR media application was rated highly.

2 Related work

Three-dimensional visual interfaces reproduce viewing objects existing in the real world. AR is defined by Azuma as systems that have the three characteristecs, (i) Combines real and virtual; (ii) Interactive in real time; (iii) Registered in 3D. In recent years, the number of use cases for AR as a medium for viewing exhibits in museums and art galleries has increased. Fenu et al. asked 34 subjects who visited the Svevo Museum autonomously with their smartphone app using AR [1]. They analyzed their behavioral records, and the items were rated highly, regarding the overall satisfaction, novelty, aesthetics of the user interface, and degree of interest for the content. Tillion et al. classified visitors' learning experiences in museums into two types, sensitive and analytical, and investigated the results of AR guides [9]. According to their results, the presentation of appropriate information by the AR guide, such as the materials of paintings and the introduction of other works, may promote the *Analytical Activity*.

In 2014, we established the SDM consortium[10] for targeting new research areas and markets involving object-based digital media and Internet-by design audio-visual environments. SDM is an architectural approach to media as a service, by the virtualization and abstraction of networked media infrastructure. LiVRation [5] was a system for interactive playback media from a free viewpoint using a head-mounted display. Web360^2 [6] was designed for viewing 3D contents on a browser with tablets, and was deployed as a WebVR application. Both applications accepted interactive manipulation from viewers, and more than half of the total number of responses were for the top two ratings combined in their subjective evaluations using a seven-point Likert scale.

3 Co-Sound

3.1 Design

We propose a platform that enables multiple people to view and manipulate the same content by playing an object-based music event using interactive AR on the web. Co-Sound satisifes the following requirements.

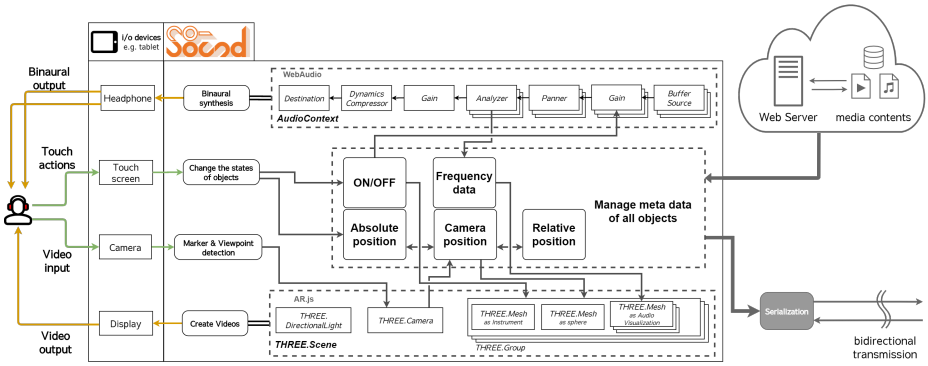


Fig. 1: Design and implementation of Co-Sound

1. Interactive viewing between viewers and contents
2. Bidirectional communication among viewers
3. Object-based structuring of media data
4. Viewing experience regardless of specific devices

Figure 1 shows an overview of Co-Sound system design and implementation. Co-Sound derives the audio data of the music event based on SDM ontology from the database, and centrally manages the displayed virtual objects. Viewers input video information and touch actions, and Co-Sound outputs binaural audio with camera images of virtual objects superimposed on them. Marker detection from the input video estimates the coordinates of the camera, and those of each virtual object are determined by referring to the position information of the recorded data. Real-time rendering of AR images and sounds in response to touch actions realizes user interactivity. Moreover, Co-Sound synchronizes the virtual space with other devices by communicating the serialized object data.

3.2 Implementation

Co-Sound was implemented using AR.js v1.5.0[1] and aframe.js v0.9.2[2], which process the marker recognition and camera location estimation. Three.js v0.110.0[3] renders AR objects and audio visualization.

Three-dimensional audio on browser was implemented using WebAudio. The nodes are chained on the **AudioContext** from the **BufferSource** node got by HTTP request to the **Destination** node. The ON/OFF operation of the sound was represented by setting the gain value of the **Gain** object, which is a gain adjustment node, to zero or a constant. Similar to $Web360^2$ [6], the visualization of the sound was represented by using the **AnalyzerNode.getByteFrequency-**

[1] https://github.com/jeromeetienne/AR.js(Accessed on 01/05/2020)
[2] https://aframe.io/blog/arjs/(Accessed on 01/05/2020)
[3] https://threejs.org/(Accessed on 01/05/2020)

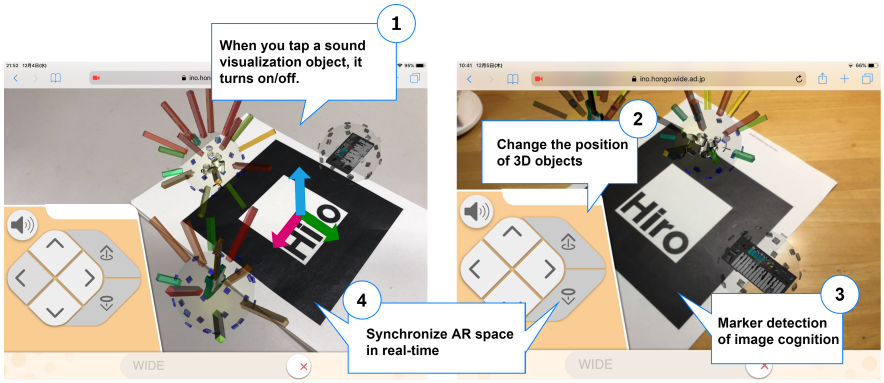


Fig. 2: Co-Sound screenshots

Data() method in WebAudio. The system obtained the frequency domain data from the time domain data and represented the effective frequency band by converting it to the length and color of the box objects.

We propose the shared and synchronized digital space with WebRTC instead of WebSocket. WebRTC is a technology of peer-to-peer (P2P) real-time connections on web browser. DataChannel, which is one of the types of WebRTC for binary data transport, adopts Stream Control Transmission Protocol (SCTP), and can ensure reliable sequential transport of messages with congestion control. Santos-González reported that its packet transmission rate is higher than Real Time Streaming Protocol [7]. We employed SkyWay v2.0.1[4], a platform as a service (PaaS) designed as a real-time interactive multimedia service. SkyWay provides a signaling server for WebRTC connections, TURN server for packet relay, and WebSocket server. These servers are publicly stated to have been located in Tokyo. Two types of communication methods and protocols were implemented for the comparison experiments: (1) mesh type connection using WebRTC, and (2) start type connection using WebSocket. The open source of SkyWay JavaScript software development kit (SDK) implements WebSocket for room-type binary data communication; for this reason, we improved it to build a mutual DataChannel connection between peers even in room type.

4 Evaluation and Discussion

4.1 Performance evaluation

In the following experiments, we evaluated the delay of AR spatial synchronization by measuring round trip time (RTT). In the field of online gaming, the QoE is closely related to the response delay [4]; hence, measuring the delay is one of the indicators to measure the QoE of the spatial synchronization function of Co-Sound.

[4] https://github.com/skyway/skyway-js-sdk(Accessed on 01/05/2020)

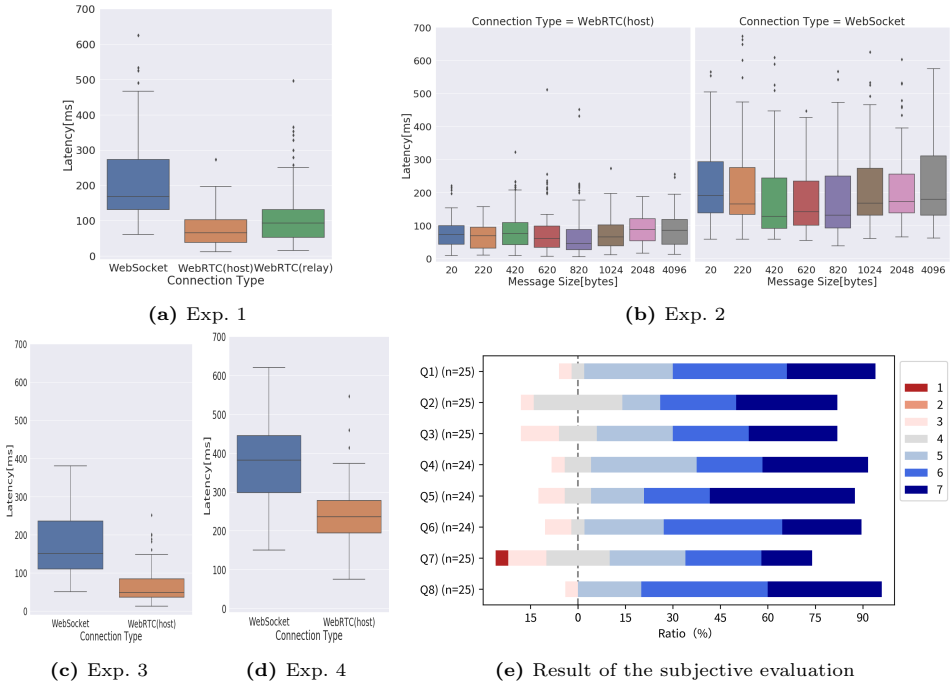
Table 1: Co-Sound measurement environments

	OS	CPU	Memory
Laptop	Windows 10 version 1809	Intel® Core™ i7-8550U	16 GB
Tablet	iOS 12.3.1	Apple A10X Fusion	4 GB
Smartphone	Android 9, EMUI version 9.1.0	HiSilicon Kirin 960	4 GB

We conducted experiments to measure the performance of Co-Sound spatial synchronization under the following four conditions. One terminal sent a test dummy file, and the other sent it straight back. We define the RTT as the time taken for a series of these transmissions. We did not consider the delay fluctuation caused by the differences in the packet processing performance of each server. Table 1 shows devices used in this experiments.

In the first experiment, we selected three types of communication protocols as those available to web browsers: (1) WebRTC in LAN (host); (2) WebRTC via TURN server (relay); and (3) WebSocket. Fig. 3a shows that the average RTT was 210 ms and 73 ms with WebSocket and WebRTC (host), respectively, which means that WebRTC was shortened by 65.0%. The average RTT with WebRTC (relay) was 107 ms. It also illustrates that the standard deviations were derived as 116 ms, 47 ms, and 87 ms, which implies that the variation in delay time was suppressed. In the second experiment, we measured RTT when various sizes of messages were transferred: 20 B, 120 B, 220 B, 420 B, 820 B, 1 KiB, 2 KiB, and 4 KiB. The result is shown in Fig. 3b. Message size had little influence on the average RTT and the standard deviation, irrespective of the protocols used. For sizes of 20 to 4096 B, the average RTT for both protocols was approximately 80 ms and 200 ms, respectively, which was constant regardless of the message size. In the third experiment, we evaluated RTT when the number of connected devices was changed. One to three smartphones shown in Table 1 joined the same room in addition to the laptop and the tablet. Fig. 3c (compared to Fig. 3a) demonstrates the result of Exp. 3. The average RTT when two and five devices joined was 65 ms and 170 ms, respectively. In the forth experiment, we measured RTT when two kinds of devices were used. The laptop shown in Table 1 and the tablet or the smartphone was used. Fig. 3d (compared to Fig. 3a) shows the result of Exp. 4. In the case of the smartphone, the average RTT was 240 ms for WebRTC and 360 ms for WebSocket. It can be inferred that the performance of the device has a significant impact on the delay, irrespective of the protocols adopted.

From Exp. 1–4, it was concluded that the proposed method employing WebRTC was more appropriate for real-time AR spatial synchronization. Although the evaluation of the QoE in spatial shared AR has not been determined yet, Nishibori's study on delay recognition in music sessions over the Internet reported that the delay is recognized at 30 ms or more, and the performance becomes difficult at 50 ms or more [12]. Vlahovic reported that the player's score and QoE decrease over 100 ms in first-person-shooting games in VR [11]. The results of these experiments show that the average delay for WebRTC communi-



(a) Exp. 1 **(b)** Exp. 2

(c) Exp. 3 **(d)** Exp. 4 **(e)** Result of the subjective evaluation

Fig. 3: (a) – (d) were the results of Exp. 1–4. RTT by WebRTC was shorter than that by WebSocket. RTT was not dependent on message sizes but the performance of devices. (e) was the result of questionnaire survey.

cation is less than 50 ms, and P2P in the same LAN could reduce the overhead by using SCTP and retain a lower latency than that using HTTP. Moreover, the transmission delay was independent of the message size and the number of devices within the range measured in the experiments. Even when the payload of AR spatial data became longer because of the increase in the number of AR objects and the complexity of the attributes, Co-Sound could be considered to be highly scalable with the real-time synchronization.

4.2 Subjective evaluation

We conducted a questionnaire survey to evaluate the QoE of Co-Sound. The survey was carried out from December 6, 2019 to December 17, 2019. Subjects were asked to experience free-viewpoint viewing, turning individual audio on and off, and moving the AR objects, and then answer the questionnaire. Responses were obtained from a total of 25 people, including 24 men and one woman. Concerning the age composition, 20 people were in their 20s, two in their 30s, one in his 40s, and two in their 50s. Apple iPad Pro (10.5 inches) iOS 12.3.1 and Sony WH-1000XM2 served as a viewing device and a headphone, respectively.

The questionnaire items were evaluated using a seven-point Likert scale, ranging from 1 to 7 (worst:1, best:7), for each of the questions Q1–Q8. The eight questions are shown in Fig. 3e.

Q1) Did you hear the sound from the direction of the AR image?
Q2) Did the sound match the distance of the AR image?
Q3) When you moved the AR image, did you feel that the sound move with it?
Q4) When you changed your viewpoint, did you feel the sound move with it?
Q5) Is it intuitive to turn on/off the audio objects using the audio visualizer?
Q6) Is it intuitive to move the AR image using the controller?
Q7) Was the recognition accuracy of the Ar markers sufficient?
Q8) Can you interact with the 3D contents on web browser?

Q1–Q4 regarded the fundamental three-dimensionality of the audio. Q5–Q6 were regarding the user interface, Q7 the accuracy of the marker detection, and Q8 the general QoE of Co-Sound.

Fig. 3e depicts the results. The vertical axis shows questions from Q1 to Q8 and the number of valid responses; the horizontal axis shows the ratio of responses for the seven-point evaluation, from 1 to 7, as a stacked bar graph. The middle of the response ratio of score 4, which represents the mid-term evaluation, was placed at the origin. The more ratings 5, 6, and 7 were given, the more the stacked bar was biased in the positive direction, and vice versa.

For all items except for Q7, the total response ratio of scores 6 and 7 was more than 50%, and as for Q8, it was 76%. On the other hand, the average rating of Q7 was 4.96, the ratio of the highest rating score was 16%, and the lowest rating score 1 was present. Q7 was the only question that had an average rating of less than 5, and the ratio with a rating of score 7 was also the lowest.

Although more than half of the responses of Q1–Q4 gave a high rating, the total response ratio of scores 5–7 in Q2 and Q3 was approximately 70%, while that in Q1 and Q4 was more than 85%. $Web360^2$[6] reported that the evaluation by the questionnaire as for the audio was dispersed, because the questions were ambiguous; for this reason, We classified the audio three-dimensionality into four types. This illustrates that the direction tracking of the audio to the AR image was excellent, but the distance tracking of the audio was not satisfactory. I would suggest that this is because a binaural algorithm employed by WebAudio PannerNode is simple and the calibration with real space is inadequate. The results of Q5, Q6, and Q8 show that the user interface of Co-Sound was rated as highly as $LiVRation$ and $Web360^2$, and the QoE of an interactive medium with AR was also high. ARToolkit, which is used in AR.js, adopts a rudimentary algorithm for marker detection and is known for its high false-negative rate [2], which appeared in the result of Q7. It can be asserted that WebAR is not accurate enough to obtain a high rating from users.

5 Conclusion

In this study, we proposed an interactive audio-visual medium using WebAR, Co-Sound. By designing a multimodal interface that dynamically renders AR

according to object operations from viewers, we presented a digital space with high affinity to the real space and interactive content viewing. Furthermore, the low-latency bidirectional communication among devices enabled users to interact with each other by allowing them to become the senders and receivers of content.

In future work, we plan the integration of real space and digital space. The current version of Co-Sound displays a music event on a marker; however, we must incorporate the advantage of AR and the induction from real to digital.

References

1. Fenu, C., Pittarello, F.: Svevo tour: The design and the experimentation of an augmented reality application for engaging visitors of a literary museum. International Journal of Human-Computer Studies **114**, 20 – 35 (2018)
2. Fiala, M.: Artag, a fiducial marker system using digital techniques. vol. 2, pp. 590 – 596 vol. 2 (07 2005)
3. ITUR Rec. Itu-r bs 2051-0 (02/2014): Advanced sound system for programme production. Int. Telecommun. Union, Geneva, Swizerland (2014)
4. Jarschel, M., Schlosser, D., Scheuring, S., Hoßfeld, T.: An Evaluation of QoE in Cloud Gaming Based on Subjective Tests. In: 2011 Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. pp. 330–335
5. Kasuya, T., Tsukada, M., Komohara, Y., Takasaka, S., Mizuno, T., Nomura, Y., Ueda, Y., Esaki, H.: Livration: Remote vr live platform with interactive 3d audio-visual service. In: IEEE Games Entertainment & Media Conference (IEEE GEM) 2019. pp. 1–7. Yale University, New Haven, CT, U.S. (2019)
6. Kato, S., Ikeda, T., Kawamorita, M., Tsukada, M., Esaki, H.: Web360^2: An Interactive Web Application for viewing 3D Audio-visual Contents. In: 17th Sound and Music Computing Conference (SMC). Torino, Italy (2020)
7. Santos-González, I., Rivero-García, A., González-Barroso, T., Molina-Gil, J., Caballero-Gil, P.: Real-Time Streaming: A Comparative Study Between RTSP and WebRTC. In: Ubiquitous Computing and Ambient Intelligence. pp. 313–325. Springer International Publishing, Cham (2016)
8. Silzle, A., Sazdov, R., Weitnauer, M.: The EU Project ORPHEUS: Object-Based Broadcasting-For Next Generation Audio Experiences. the 29th Tonmeistertagung -VDT International Convention (01 2016)
9. Tillon, A.B., Marchal, I., Houlier, P.: Mobile augmented reality in the museum: Can a lace-like technology take you closer to works of art? In: 2011 IEEE International Symposium on Mixed and Augmented Reality - Arts, Media, and Humanities. pp. 41–47 (Oct 2011)
10. Tsukada, M., Ogawa, K., Ikeda, M., Sone, T., Niwa, K., Saito, S., Kasuya, T., Sunahara, H., Esaki, H.: Software defined media: Virtualization of audio-visual services. In: 2017 IEEE International Conference on Communications (ICC). pp. 1–7 (2017)
11. Vlahovic, S., Suznjevic, M., Skorin-Kapov, L.: Challenges in Assessing Network Latency Impact on QoE and In-Game Performance in VR First Person Shooter Games. In: 2019 15th International Conference on Telecommunications (ConTEL). pp. 1–8 (July 2019)
12. Yu Nishibori and Yukio Tada and Takuro Sone: Study and Experiment of Recognition of the Delay in Musical Performance with Delay. IPSJ SIG Technical Reports **53**, 37–42 (dec 2003)