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}@hongo.wide.ad.jp, hiroshi@wide.ad.jp

Abstract. An Internet-based media service platform can control recording processes and manage video and audio data interconnected by an IP network. 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 that allows us to interact with both elements in real space and digital space information. 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 an interactive audio-visual playback application for music events, using WebAR. co-Sound was designed as a multimodal interface that dynamically renders object-based AR in response to various actions from viewers on a web browser with low entry costs. Furthermore, by sharing AR objects among multiple devices in real time and bidirectionally, the relationship between users and contents was extended, and interaction among multiple users in the same AR space was possible. We implemented a prototype application, measured the performance of the AR spatial synchronization, and conducted a questionnaire-based evaluation. For subjective evaluation, 25 people experienced co-Sound and completed a questionnaire. We confirmed that the system was developed by object-based method with AR, achieved low-latency synchronization to accept operations from multiple users in real time, and the general acceptance of the system was very high.

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

1 Introduction

With the spread of the high-capacity communication environments, the emergence of the fifth-generation mobile communication system, also known as 5G, and the next-generation wireless communication standard, IEEE802.11ax (WiFi-6), and development of audio lossless delivery technologies, such as 4K video and 360° video, has become more prevalent than ever. However, the video content mentioned above typically remains a static viewing experience.

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. Video and audio contents distributed via the Internet, particularly contents that reproduce the objects in the real world using a three-dimensional spatial reproduction method, have been gaining in popularity.

Sound recording and playback systems can be broadly divided into three categories [6]. Object-based audio (OBA) has the following characteristics [12].

1. Multiple objects that exist in multi-dimensional space;
2. Interactive reproduction personalized to users;
3. 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 with high order ambisonics (HOA), 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, and several experiments to measure its performance and quality of experience (QoE). Such a audio-visual media platform is ideal for reproducing software-managed object audio. Furthermore, unlike Virtual Reality (VR), which reproduces actual live musical performances as it is with 360° video, AR display can freely place audio data and 3D avatars tied to it in the real world, allowing for flexible operations. By measuring the real-time response of multiple people to the system and the QoE 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.

The remainder of this paper is organized as follows. First, we discuss related work, including the QoE of OBA, use cases of AR, and software defined media (SDM) consortium. Subsequently, in Sect. 3, we organize the requirements and in Sect. 4, we describe the design and implementation of the proposed method, co-Sound. Sect. 5 details the experimental setup along with the experimental results and Sect. 6 concludes the paper.

2 Related work

2.1 Object-based audio

OBA provides novel viewing experience and creates interactive, personalized, scalable, and immersive content. In ORPHEUS, the European Commission HORIZON 2020 research project, the object-based radio reproduction app was evaluated through two QoE tests [13]. It was reported that the general acceptance of the new features and functions that came with OBA was significantly high, and the usability was also rated highly.

2.2 Augmented Reality (AR)

Three-dimensional visual user interfaces reproduce viewing objects existing in the real world. AR is defined by Azuma [3]. In this study, we focus on the display functions of AR.

1. Combines real and virtual
2. Interactive in real time
3. 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 to view the exhibits autonomously with their smartphone app using AR [4]. They analyzed their behavioral records and the five-point ratings for the app, and the items were rated highly, regarding the overall satisfaction, novelty, aesthetics of the user interface, and degree of interest for the content; this indicates that viewing the exhibits with AR was useful. Tillion et al. classified visitors' learning experiences in museums into two types, sensitive and analytical, and investigated the results of AR guides on art appreciation [14]. According to their results, the AR guide may interfere with the sense of immersion in a work of art, thereby having a negative impact on *Sensitive Activity*, but the presentation of appropriate information, such as the materials of paintings and the introduction of other works, may promote the *Analytical Activity*.

AR attracts attention as an interface and medium for computer-supported cooperative work. Lukosch asserted that communicating via AR in some way had the potential to ameliorate cooperative work. The first cooperative AR was Studierstube, and was proposed by Schmalstieg [11]. He reported that cooperative AR did not interrupt natural communication, such as voice and gestures, because there were minimum virtual contents, unlike VR. In 2019, applications such as Cloud Anchors[1] and Azure Spatial Anchors[2] emerged and enabled the same AR content to be viewed across multiple devices by sharing and storing spatial recognition information. Particularly, a host device saves some 3D maps in the cloud database, and the other devices get the 3D spatial information to synchronize the AR.

2.3 Software defined media

In 2014, we established the SDM consortium[3] 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,

[1] https://developers.google.com/ar/develop/java/cloud-anchors/overview-android(Accessed on 01/05/2020)

[2] https://azure.microsoft.com/ja-jp/services/spatial-anchors/(Accessed on 01/05/2020)

[3] https://sdm.wide.ad.jp/

by the virtualization and abstraction of networked media infrastructure. SDM construct their original architecture [15] and have the following goals.

1. Software-programmable environment of 3D audio-visual services
2. Mixing 3D audio objects from multiple sources
3. Augmented sound and video effects via software rendering
4. User interaction

LiVRation [8] was a system for interactive playback media from a free viewpoint using a head-mounted display. This system reproduced the actual music event in a virtual space, which was recorded by 360-degree cameras from multiple locations and directional/omnidirectional microphones by each instrument. Web360^2 [9] 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 Objectives and Requirements

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. For this proposal, the following requirements are given.

Interactive viewing between viewers and contents We deploy recorded music events on AR. AR will make it possible to project a digital space in accordance with the real space in contrast to VR, which divides two spaces ultimately. Furthermore, the system will allow access to individual audio-visual source data in response to the viewer's actions. It will realize dynamic playback and manipulating media of the viewer's own, such as selecting and rearranging audio objects that cannot be performed by conventional static content playback. Audio objects as well as visual ones will be presented in a three-dimensional acoustic format that is expected to give a sense of immersion and presence following the viewer's movements.

Bidirectional communication among viewers Sharing and synchronizing the digital space among multiple viewers is expected to create a new kind of interactivity that is different from regular content playback, which is limited to a one-to-one relationship between a viewer and a content. As the system will store and manage object-based media data, a viewer as a content receiver will be able to serve as a provider. Media must be able to accept operations from multiple people and present them in real time rather than being a one-way playback device.

Object-based structuring of media data AR contents are structured into object-based media data. By storing and managing the visual, audio, and AR data of target music events based on the SDM architecture, it is possible to develop a flexible playback environment that cannot be realized by conventional methods, such as channel-based and HOA/ambisonics.

Viewing experience regardless of specific devices The proposed method is implemented as a WebAR application. Table 1 shows the current AR development environments. WebAR has a significantly lower cost of entry, because it does not require installing specific applications and is promptly accessible as soon as users open a web browser. Browser kernel-based, specific application-based, and hardware-based AR are inferior in that they cannot be cross-platformed. A single platform, which meets the requirements of a particular operating system (OS), will make developers and users spend high-cost unifying apps between OS; therefore, we adopt web application based on Pure frontend.

Table 1: Comparing AR development environments

Type	**device**	**entry cost**	**responsiveness**	**platform**
Pure front end	web browser	high	low	cross platform
Browser kernel	browser kernel	high	middle	browser
Application	smartphone or tablet	high	middle	OS
Hardware	head-mounted display	low	high	hardware

4 co-Sound

4.1 Recording dataset

The music event used in the co-Sound application was a recording of the Musilogue Band's concert held at Billboard Live Tokyo in Roppongi Midtown on January 26, 2017 [16]. In that concert, the band was composed of three types of instruments: drums, electric bass, and keyboard. The microphones were set up for each instrument, and the sound source was recorded individually. In this study, **Target** information of the data structure of musical events defined in SDM ontology [2] was applied. We used the attributes, position information, and audio information of each instrument in the concert described above.

4.2 Design overview

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

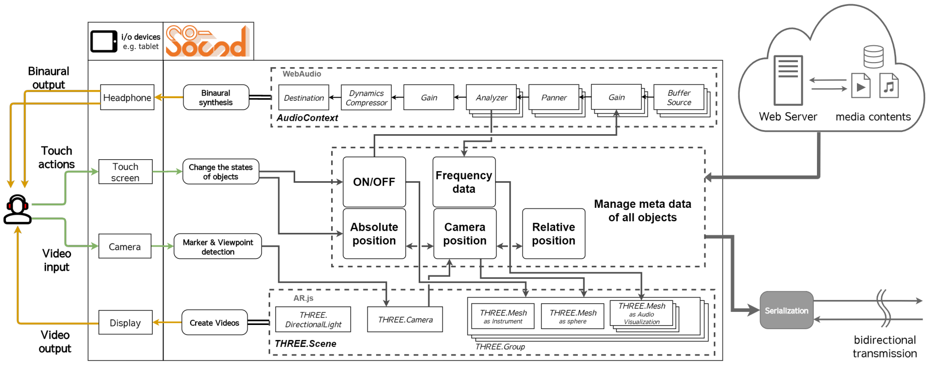


Fig. 1: Design and implementation of co-Sound

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.

co-Sound synchronizes the virtual space with other devices by communicating the serialized object data that are managed by co-Sound.

4.3 Implementation

Overview co-Sound is an interactive medium with AR based on the design approach above, and reproduces a music event recorded by SDM on AR. The application data are displayed on a web browser, and the media data of each instrument, including audio files, position data, and 3D model files, are sent from a web server simultaneously. Table 2, Table 3, and Fig. 2 illustrate the used framework, devices, and screenshots of co-Sound, respectively.

Table 2: co-Sound implemented environments

WebAR framework	AR.js v1.5.0, aframe.js v0.9.2
WebGL framework	Three.js v0.110.0
Browser	Chrome v79, Safari v604.1

WebAR AR.js[4] and aframe.js[5] process the marker recognition and camera location estimation. Three.js[6] renders AR objects and audio visualization. Rendered

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

Table 3: co-Sound measurement environments

	OS	CPU	Memory
Laptop	Windows 10 version 1809	Intel® Core^{TM} 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

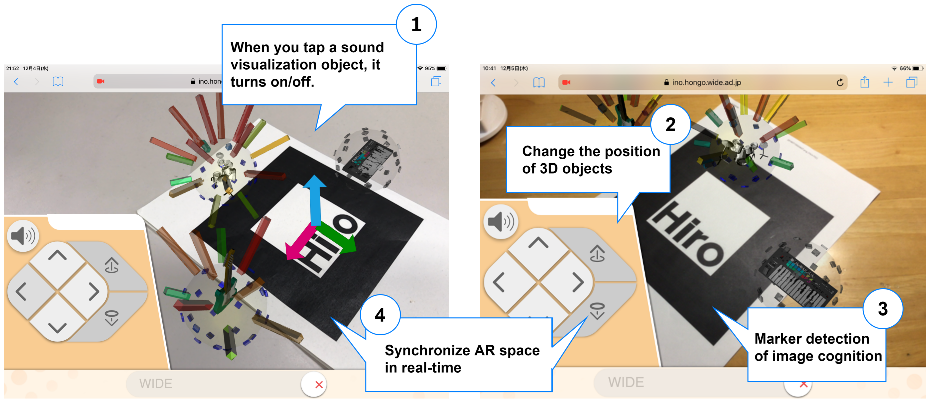

1
When you tap a sound visualization object, it turns on/off.
2
Change the position of 3D objects
3
Marker detection of image cognition
4
Synchronize AR space in real-time
Hiro
Hiro
WIDE
WIDE
ino.hongo.wide.ad.jp

Fig. 2: co-Sound screenshots

groups are divided by the instrument and defined by **THREE.Group()** class in AR.js, which enables AR objects to be processed in the same way as the object-based audio in SDM architecture is.

User interactions A user inputs actions using a smartphone or a tablet, and co-Sound updates the content in response to the viewer's actions. The user can change the volume of instruments by touching the object and also arrange their position by the cross key controller.

3D sound effect Three-dimensional audio on browser was implemented using WebAudio[7]. 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 $Web360^2$ [9], the visualization of the sound was represented by using the **AnalyzerNode.getByteFrequencyData()** 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.

[7] https://www.w3.org/TR/webaudio/(Accessed on 01/05/2020)

Connection between devices We propose the shared and synchronized digital space with WebRTC[8] 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. In general, the transmission is slow because of the overhead in these processes when there is a considerable amount of packet loss in the reliable mode. However, WebRTC adopts SCTP over DTLS over UDP for NAT traversal, and Santos-González reported that its packet transmission rate is higher than Real Time Streaming Protocol [10].

We employed SkyWay v2.0.1[9], 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. In this study, co-Sound was designed to create a room divided by namespace and peers to establish connections with each other in the room. 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 the room type.

5 Evaluation and result

5.1 Performance evaluation

The nth peer is referred to as P_n. In the following experiments, we evaluated the delay of AR spatial synchronization by measuring the round trip time (RTT). In the field of online gaming, the QoE is closely related to the response delay [7]; hence, measuring the delay is one of the indicators to measure the QoE of the spatial synchronization function of co-Sound. Fig. 3 shows the network topology. We express t_1^A as the internal processing time from the moment the P_1 browser issues the transmission instruction to the moment it is sent through the user space, kernel space, and network interface card (NIC) to the Internet, and express t_1^B as the transport time from P_1 to P_2 through a local area network (LAN) or a wide area network (WAN). P_2 passes the packet received from P_1 to the browser and sends the same packet back to P_1. We express t_2^P as the processing time for sending and receiving a packet in a browser. From the above, the time of a series of processes can be expressed as in Equation 1.

$$\mathrm{RTT} = 2 \times (t_1^A + t_{12}^B + t_2^A) - t_2^P \tag{1}$$

[8] https://www.w3.org/TR/webrtc/(Accessed on 01/05/2020)

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

In this measurement, we did not consider the delay fluctuation caused by the differences in the packet processing performance of each server.

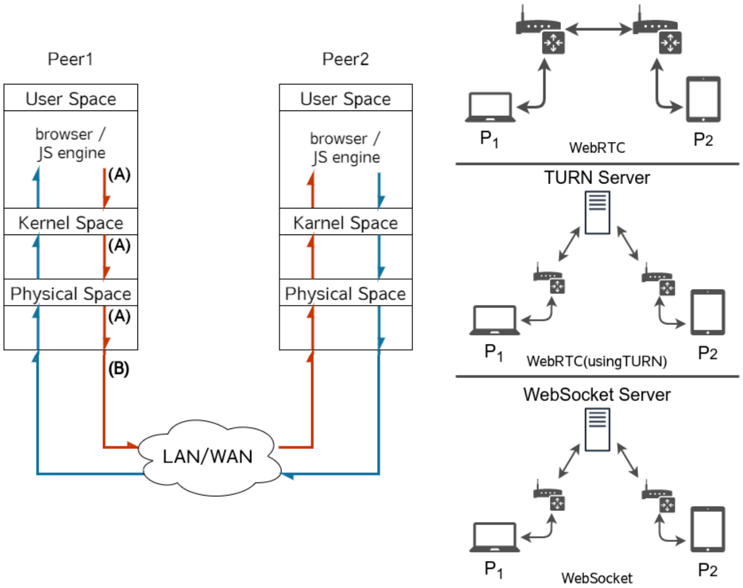


Fig. 3: Network topology

Exp. 1: Delay dependent on the network protocols We show that co-Sound can be synchronized with a lower latency using a parameter of the protocol for P2P communication of AR object data. For comparison, 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. We measured RTT when 1 KiB JSON data were transferred 100 times at 5-second intervals, using a laptop as P_1 and a tablet as P_2.

Fig. 4a 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 can be inferred that the proposed method of WebRTC can transfer packets with a lower delay than that of WebSocket even when packets are relayed; furthermore, the standard deviations were derived as 116 ms, 47 ms, and 87 ms, which implies that the variation in delay time was suppressed.

Exp. 2: Delay dependent on the message size 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 rest of the conditions were the same as in Exp. 1.

The result is shown in Fig. 4b. 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 WebRTC and WebSocket was approximately 80 ms and 200 ms, respectively, which was constant regardless of the message size.

Exp. 3: Delay dependent on the number of devices We evaluated RTT when the number of connected devices was changed. One to three smartphones shown in Table 3 joined the same room as P_3–P_5 in addition to the laptop and the tablet. The rest of the conditions were the same as in Exp. 1.

Fig. 4c (compared to Fig. 4a) demonstrates the result of Exp. 3. The average RTT when two and five devices joined was 65 ms and 170 ms, respectively.

Exp. 4: Delay dependent on the performance of devices We measured RTT when two kinds of devices were used. The laptop shown in Table 3 served as P_1 and either the tablet or the smartphone was used as P_2. The rest of the conditions were the same as in Exp. 1.

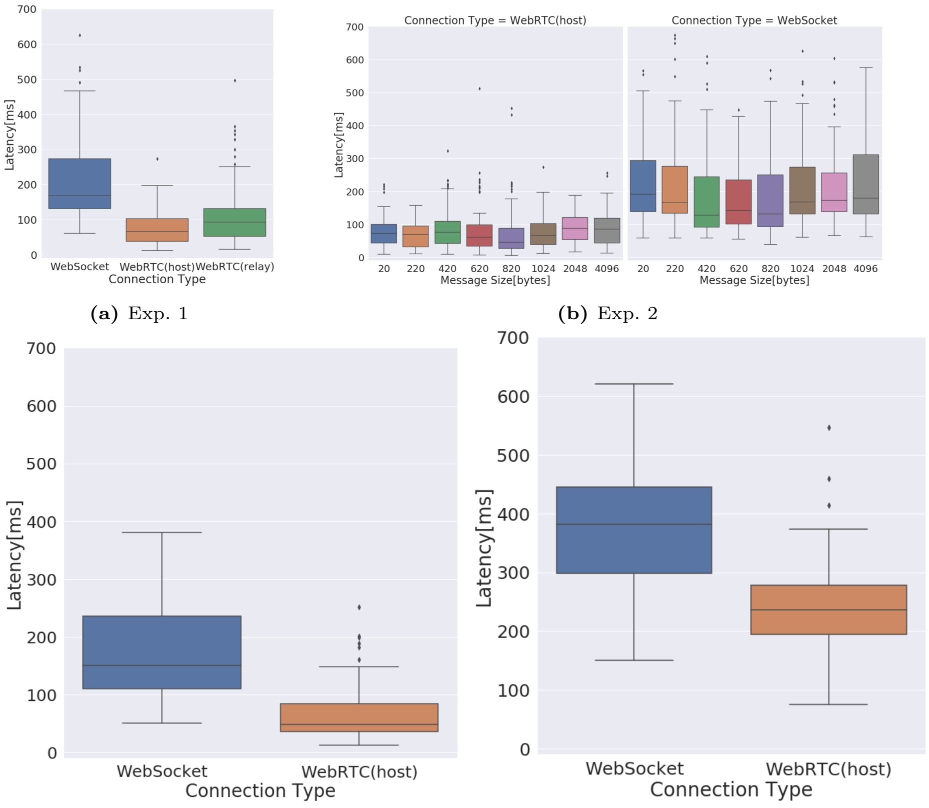
Fig. 4d (compared to Fig. 4a) shows the result of Exp. 4. In the case of the smartphone, the average RTT was 240 ms for WebRTC and 360 ms for WebSocket. It can be inferred that the performance of the device has a significant impact on the delay, irrespective of the protocols adopted.

5.2 Subjective evaluation

Method and subject We conducted a questionnaire survey to evaluate the viewing experience of interactive three-dimensional content using co-Sound. The survey was carried out from December 6, 2019 to December 17, 2019. Initially, the usage of co-Sound was explained; subsequently, an experienced person was permitted to operate a device freely, after which data were acquired through a 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.

Questionnaire items 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. 5.

Questions Q1–Q4 regarded the fundamental three-dimensionality of the audio. $Web360^2$[9] 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. Questions Q5–Q6 were regarding the user interface, Q7 the accuracy of the marker detection, and Q8 the general QoE of co-Sound.



(a) Exp. 1

(b) Exp. 2

(c) Exp. 3

(d) Exp. 4

Fig. 4: Exp. 1–4 results

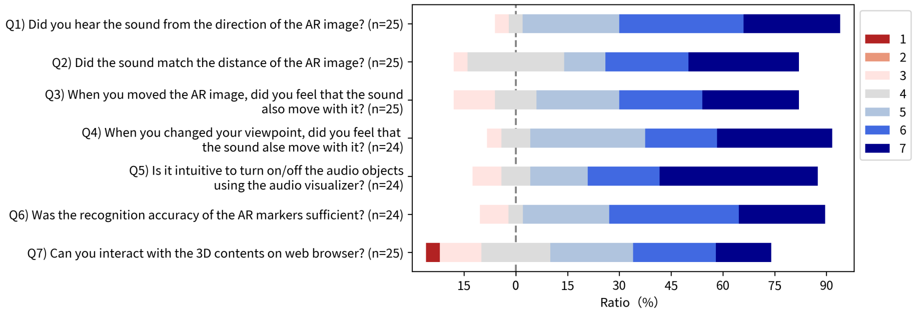


Fig. 5: Eight questions results

Result Fig. 5 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 score 7, which is the highest rating, 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.

5.3 Discussion

Performance of co-Sound 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 [18]. Vlahovic reported that the player's score and QoE decrease over 100 ms in first-person-shooting games in VR [17]. The results of these experiments show that the average delay for WebRTC communication 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.

Questionnaires Although more than half of the responses of Q1–Q4, which regarded three-dimensional audio, 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%. 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 [5], which appeared in the result of Q7. It can be asserted that WebAR is not accurate enough to obtain a high rating from users.

6 Conclusion

In this study, we proposed an interactive audio-visual medium using WebAR, and developed a web application, co-Sound. By designing a multimodal interface that dynamically rendered 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 between viewing devices enabled users to interact with each other by allowing them to become the senders and receivers of content. We conducted two experiments to verify these proposals. The first one was to measure RTT as the performance of co-Sound. The delay of spatial synchronization using WebRTC was approximately 50 ms, both in the same space and in remote places. The second one was to evaluate the QoE. While we recognized the importance of marker-detection accuracy, more than half of the subjects rated highly and clarified the dispersion in the evaluation of audio, which was attributed not to the sense of direction but to the sense of distance.

In future work, we plan two improvements. The first one is 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 space to digital space. The second one is the improvement of marker-detection accuracy. As shown in the results of this study, the accuracy of detection lowers the QoE. As the available resources for browser kernel-based AR and application-based AR are limited, the utilization of mobile edge computing (MEC) for AR has been investigated [1]. MEC will also be applied to WebAR, where the operation processing for AR currently performed on mobile devices can be distributed to external resources, and high-precision tracking and real-time rendering can be expected.

References

1. Al-Shuwaili, A., Simeone, O.: Energy-Efficient Resource Allocation for Mobile Edge Computing-Based Augmented Reality Applications. IEEE Wireless Communications Letters **6**(3), 398–401 (1 2017). https://doi.org/10.1109/LWC.2017.2696539
2. Atarashi, R., Sone, T., Komohara, Y., Tsukada, M., Kasuya, T., Okumura, H., Ikeda, M., Esaki, H.: The Software Defined Media Ontology for Music Events. In: Workshop on Semantic Applications for Audio and Music (SAAM) held in conjunction with ISWC 2018. pp. 15–23. Monterey, California, USA. (2018)
3. Azuma, R.T.: A survey of augmented reality. Presence: Teleoperators and Virtual Environments **6**(4), 355–385 (Aug 1997)
4. 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). https://doi.org/https://doi.org/10.1016/j.ijhcs.2018.01.009, advanced User Interfaces for Cultural Heritage
5. Fiala, M.: Artag, a fiducial marker system using digital techniques. vol. 2, pp. 590 – 596 vol. 2 (07 2005). https://doi.org/10.1109/CVPR.2005.74

6. ITUR Rec. Itu-r bs 2051-0 (02/2014): Advanced sound system for programme production. Int. Telecommun. Union, Geneva, Swizerland (2014)
7. 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 (2011)
8. 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). https://doi.org/10.1109/GEM.2019.8811549
9. 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)
10. 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: García, C.R., Caballero-Gil, P., Burmester, M., Quesada-Arencibia, A. (eds.) Ubiquitous Computing and Ambient Intelligence. pp. 313–325. Springer International Publishing, Cham (2016)
11. Schmalstieg, D., Fuhrmann, A., Hesina, G., Szalavári, Z., Encarnação, L., Gervautz, M., Purgathofer, W.: The studierstube augmented reality project. Presence Teleoperators & Virtual Environments **11** (01 2003). https://doi.org/10.1162/105474602317343640
12. 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)
13. Silzle, A., Schmidt, R., Bleisteiner, W., Epain, N., Ragot, M.: Quality of experience tests of an object-based radio reproduction app on a mobile device. J. Audio Eng. Soc. **67**(7/8), 568–583 (Aug 2019)
14. 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). https://doi.org/10.1109/ISMAR-AMH.2011.6093655
15. 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 (May 2017). https://doi.org/10.1109/ICC.2017.7996610
16. Tsukada, M., Komohara, Y., Kasuya, T., Nii, H., Takasaka, S., Ogawa, K., Esaki, H.: $SDM360^2$: An Interactive 3D Audio-visual Service with a Free-view-listen Point. Transactions of Information Processing Society of Japan Digital Contents (DCON) (in Japanese) **6**(2), 10–23 (aug 2018)
17. 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). https://doi.org/10.1109/ConTEL.2019.8848531
18. 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)