Harmonic Reproduction Technology between a Real Venue and Remote Audiences

Takayuki Kurozumi, Keisuke Hasegawa, Eiichiro Matsumoto, Toshihiko Eura, and Shinji Fukatsu

Abstract

NTT Human Informatics Laboratories has been researching and developing technology to reproduce the appearance of an audience enjoying a live-streamed event online remotely from home (remote audience) at the venue where the event is being held (real venue) while harmonizing it with the situation at the real venue. At the 34th Mynavi TOKYO GIRLS COLLECTION 2022 SPRING/SUMMER held on March 21, 2022, a demonstration experiment was conducted to support the excitement of the event by using low-latency video communication and cross-modal sound search to reproduce pseudo cheers at the real venue for both real-venue and remote audience members who could not cheer due to the COVID-19 pandemic. This article introduces the activities of this demonstration experiment.

Keywords: two-way video communication, remote viewing, harmonic reproduction

1. The necessity of harmonic reproduction of video and sound

NTT has been engaged in research and development of interactive video communications that interconnect multiple remote-viewing environments and deliver high-definition video with low latency with the primary focus on delivering highly realistic video. In the Real-time Remote Cheering for Marathon Project [1], ultralow-latency communication technology with uncompressed transmission and low-latency media-processing technology were used to connect the marathon course in Sapporo, Hokkaido with the cheering venue in Tokyo in real time. This enabled spectators to send their support to the athletes from remote locations, creating a sense of presence similar to that of cheering along the course and a sense of unity between athletes and spectators, thus enabling a new way to watch the race.

To develop this initiative for home users, NTT began research on bidirectional video and audio communication between the venue of a live-streaming event (real venue) and home environment. However, there are inconveniences that arise when trying to achieve bidirectional, highly realistic video communication between the home environment and a real venue. For example, in situations such as web conferencing using video and sound between a remote-working home environment and the workplace, the background of the camera image from the home environment may show a room with a lived-in feel, or a family member's voice may be mixed in with the microphone audio, which can be awkward. In situations such as a live sporting or entertainment event, the audience members who participate remotely (remote audience) want to be present in the video to share the excitement with the audience at the real venue (venue audience) but would like to avoid information they do not want seen or heard from being distributed to the venue and other remote audience. Therefore, it is necessary to suppress unnecessary information and reproduce information that is desired to be reproduced at the real venue with a high sense of presence and harmony.

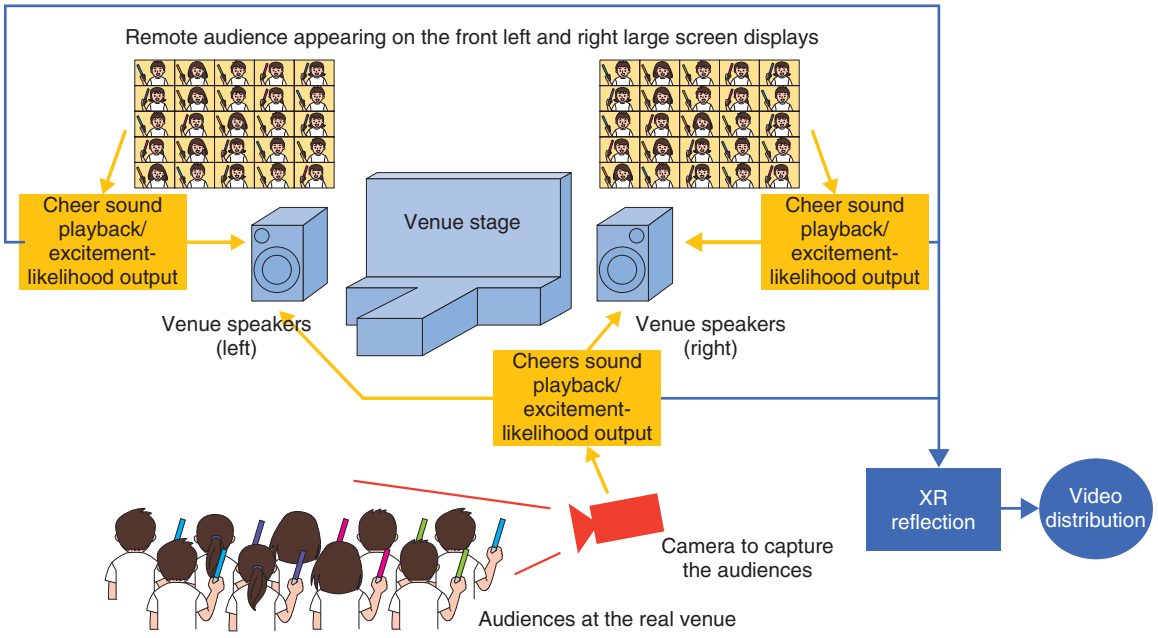


Fig. 1. Conceptual diagram of the demonstration experiment.

The real venue may also require harmonious reproduction. The COVID-19 pandemic has restricted people's activities, and even at live concerts, audiences are prohibited from cheering, even if they are wearing masks, to prevent infection. It is an uncomfortable situation for the audience at live music concerts since they cannot cheer for the popular songs that drew cheers before the pandemic, even though there are many people in attendance. It is also difficult for the performers to interact with the audience if they cannot hear the cheers, because it is difficult to understand the audience's reactions. Therefore, we studied the possibility of reproducing the harmony of images and sound so that the venue audience can feel the same excitement from the cheers as before the pandemic.

2. Joint experiment to enable two-way video communications for the home

IMAGICA EEX, NTT Communications, and NTT conducted a joint experiment to enable interactive high-resolution video viewing for home users at the 34th Mynavi TOKYO GIRLS COLLECTION 2022 SPRING/SUMMER [2] held on March 21, 2022. The experiment aimed to support the excitement of the venue audience, who were unable to cheer due to the COVID-19 pandemic, and create a sense of participation for the remote audience through interaction with the real venue. We constructed a system to reproduce cheer sounds in accordance with the excitement of the venue and remote audiences using low-latency video communication technology and cross-modal sound retrieval technology [3] and verified the reproduction of harmony.

Figure 1 shows a conceptual diagram of the entire experiment. Remote audience members remotely participated via NTT Communications' two-way low-latency communication systems (Smart vLive® [4] and ECLWebRTC SkyWay [5]) using a personal computer (PC) with a camera and appeared on the left and right large-screen displays in front of the stage at the real venue. The cheer sounds were estimated from the left and right images of the remote audience members, as described below, and the corresponding left and right speakers played the cheers in response to the excitement of the event. For the venue audience, the cheer sounds were estimated from the images taken by the camera aimed at the audience seats and played from the venue speakers. The audience could control the cheer sounds so that the cheers became louder when the audience shook their penlights faster and quieter when they shook them slower. The volume of excitement was reflected in the cheer sounds as well as in the extended reality (XR) expression of the live-streamed video, with IMAGICA EEX

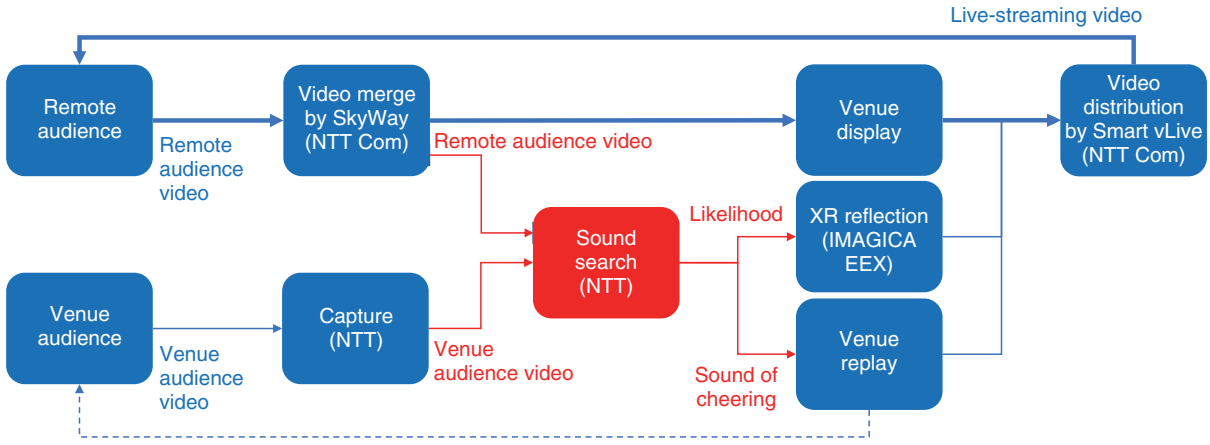


Fig. 2. System configuration and flow of processing and information.

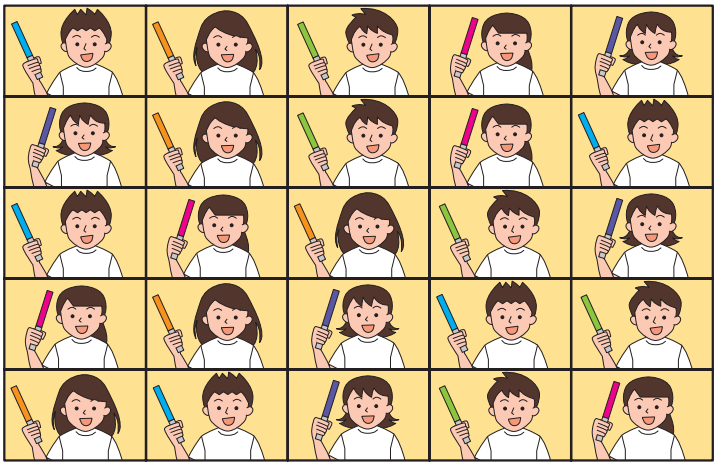


Fig. 3. Videos of remote audience waving penlights are arranged in a grid and aggregated.

creating a production in which the amount of light particles changes in accordance with the volume of excitement from the venue audience and remote audience members in the streamed video, allowing the remote audience members to enjoy the excitement of both audiences with sound and images.

2.1 System configuration

The overall system configuration is shown in **Fig. 2**. The experimental system consists of a function that lays out images of remote audience members in a tiled format, one that displays the images on a large display at the real venue, one that searches for and plays the cheer sounds from the tiled images and audience images at the real venue, one that expresses the search results in XR, and one that distributes images of the real venue reflecting these results.

The sound-retrieval system used in this project is based on machine learning technology using NTT Communication Science Laboratories' cross-modal sound retrieval technology [3] to estimate the cheer sounds from video images of audience members waving penlights. To map the images of the audience waving penlights to the cheer sounds, training data were prepared in advance by pairing the images of the venue and remote audiences waving penlights with the cheer sounds, and a model for estimating the sound from the images was trained. For the remote audience, we input aggregated video images laid out in a 5 × 5 tiled pattern, as shown in **Fig. 3**, and

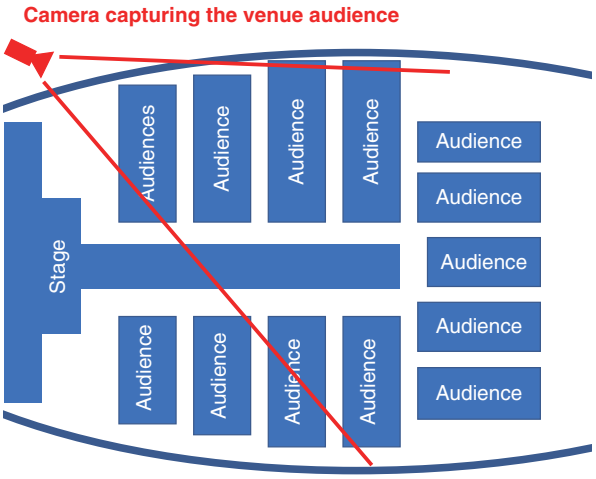


Fig. 4. Arrangement of cameras capturing the venue audience.

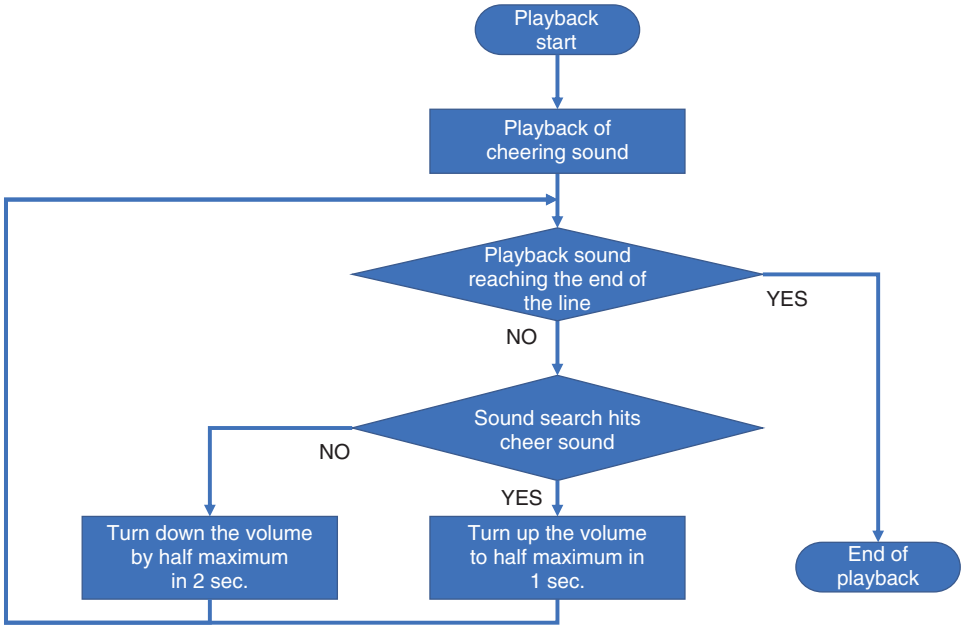


Fig. 5. Flowchart for volume determination.

retrieved and played back the corresponding cheer sounds on the basis of the penlight waving. For the venue audience, as shown in **Fig. 4**, a camera was set up in the venue to take a video of the audience, and the corresponding cheer sounds were retrieved and played back on the basis of the images of the audience waving penlights in the same way. The sound source for playback was a pre-recorded cheer sound.

In addition to searching for the cheers using the cross-modal sound retrieval technology, a method of determining the volume of the cheers using the flowchart shown in **Fig. 5** was implemented to smoothly change the volume of the cheers. An example of a hit point in the cheer-sound search and corresponding volume change is shown in **Fig. 6**. With this mechanism, it is possible to control the volume so that it increases when the penlights are continuously waved and decreases when they stop waving, thus enabling

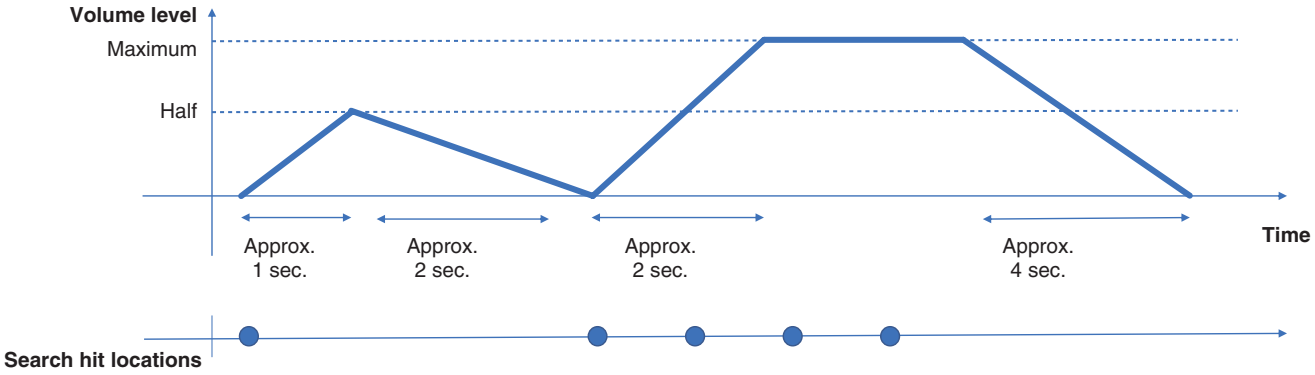


Fig. 6. Examples of search-hit locations and volume changes.

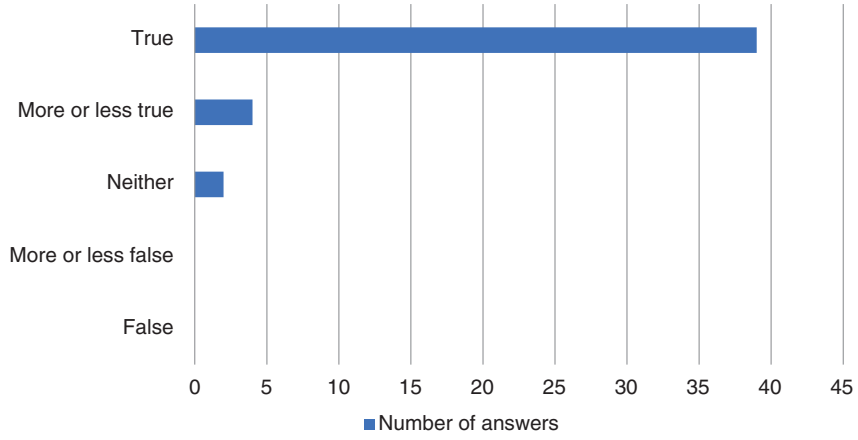


Fig. 7. Questionnaire item "I felt that it would be better to have a mechanism to produce cheers compared with the usual delivery of just watching the show" (45 respondents).

intuitive volume changes.

2.2 Evaluation by questionnaire to participants

To confirm the effectiveness of the harmonic reproduction, in which the cheer sounds are estimated from the audience images and played back at the venue, a questionnaire was sent to those participated as remote audience members during the experiment. In response to the statement, "I felt that it would be better to have a mechanism to produce cheers compared to the usual delivery of just watching the show," 86.6% of the participants responded positively on a 5-point scale, i.e., "true," "more or less true," "neither," "more or less false," and "false" (**Fig. 7**). Thus, the majority of participants had a favorable view of the harmonization reproduction system, confirming the effectiveness of delivering responses from the remote audience to the real venue.

3. Future developments

Remote audience members were asked to connect one by one from their homes, and a survey was conducted to determine what type of viewing environment they would prefer for remote participation, i.e., "In which of the following viewing situations would you prefer to watch a live webcast remotely?". More than half (68.8%) of the respondents answered, "Gather at home or a friend's house and participate with a friend from a single smartphone, PC, or monitor" (**Fig. 8**). This result may suggest that a viewing style in which good friends gather to participate in a remote-viewing environment and many such viewing environments are connected to the real venue to

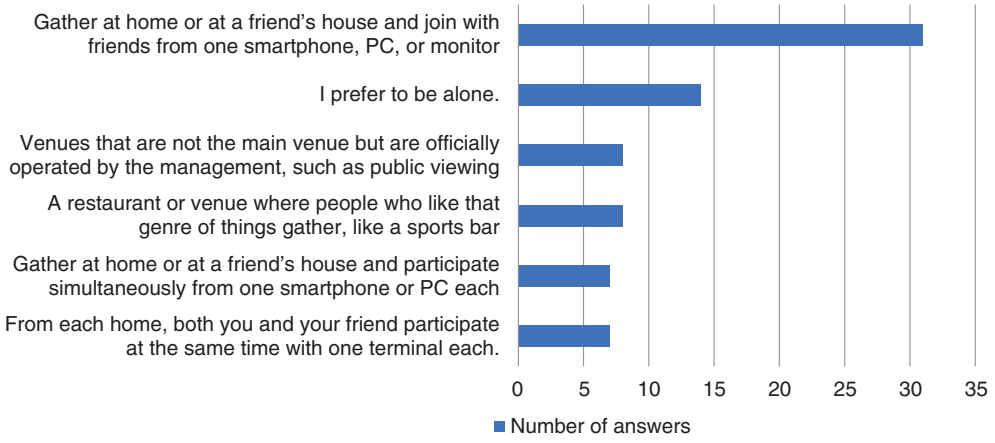


Fig. 8. Questionnaire item "Which of the following viewing situations would you prefer for watching a live webcast remotely?" (45 respondents, multiple responses allowed).

watch together will be preferred in the future. In a well-developed environment, such as a public-viewing event, there may be no problem in pursuing a high level of realism and rich transmission, including the atmosphere of the venue. However, for home-use, as mentioned earlier, if the video images captured with cameras in the home environment and the sound captured with microphones are transmitted and played back at the venue as they are, problems are expected to arise in terms of production. Therefore, NTT will continue researching and developing two-way video communications that can reproduce images and sounds in harmony so as not to interfere with live transmission by selecting information in the pursuit of reality as well as consideration of what information is to be emphasized or suppressed.

References

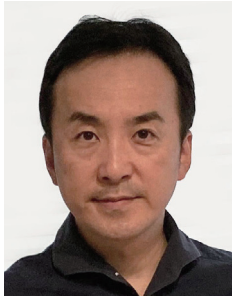
[1] S. Usui, S. Fukatsu, E. Matsumoto, M. Imoto, D. Shirai, and S. Kinoshita, "Marathon × Ultra-low-latency Communication Technology," NTT Technical Review, Vol. 19, No. 12, pp. 78–84, 2021. https://doi.org/10.53829/ntr202112fa10
[2] Mynavi TOKYO GIRLS COLLECTION 2022 SPRING/SUMMER (in Japanese), https://tgc.girlswalker.com/22ss/
[3] M. Yasuda, Y. Ohishi, Y. Koizumi, and N. Harada, "Crossmodal Sound Retrieval Based on Specific Target Co-occurrence Denoted with Weak Labels," Proc. of INTERSPEECH 2020, pp. 1446–1450, Virtual, Oct. 2020.
[4] Smart vLive® (in Japanese), https://www.ntt.com/business/services/voice-visual-communication/business-support/smartvlive.html
[5] ECLWebRTC, https://webrtc.ecl.ntt.com/en/

Takayuki Kurozumi
Senior Research Engineer, Cyber-World Laboratory, NTT Human Informatics Laboratories.

He received a B.S. in physics from the Tokyo Metropolitan University in 1997 and M.S. and Ph.D. in information science from the Japan Advanced Institute of Science and Technology, Ishikawa, in 1999 and 2007. In 1999, he joined NTT Communication Science Labs, where he was engaged in research and development (R&D) on multimedia information retrieval. From 2015 to 2017, he was with NTT Media Intelligence Labs, where he researched and developed image retrieval technology. He was with NTT DATA Corporation between 2017 and 2019, where he engaged in the planning and product development of multimedia information retrieval systems. From 2019 to 2021, he was with NTT Media Intelligence Labs, where he researched and developed compressed sensing and modeling technology. Since 2021, he has been with NTT Human Informatics Labs, where he is researching and developing distributed ultra-reality viewing technology. He was awarded the IEEE Transactions on Multimedia Paper Award in 2004.

Keisuke Hasegawa
Research Engineer, NTT Human Informatics Laboratories.

He received a B.E. and M.E. in informatics from Kyoto University in 2012 and 2014. He joined NTT WEST in 2014 and engaged in the maintenance of network facilities. Since joining NTT Service Evolution Laboratories in 2016, he has been engaged in research of media processing for the super realistic telecommunication technology Kirari!. He is a member of the institute of image information and television engineers (ITE).



Eiichiro Matsumoto
Senior Research Engineer, Cyber-World Laboratory, NTT Human Informatics Laboratories.

He received a B.E. in science engineering (information science) from Meiji University in 1999 and joined NTT the same year. From August 1999 to 2008, he was engaged in the development of geographic information systems and voice/video communication systems at NTT EAST. From 2008 to 2010, he was involved in the R&D of Internet Protocol television (IPTV)/mobile broadcasting standards and the formulation of technical standards at NTT. From 2010 to 2020, he was engaged in the planning and development of network services and systems for video distribution systems, such as IPTV and radio-frequency TV and researching video-transmission-system technologies for adding high value to relay networks and access networks at NTT EAST. He has been at his current position since 2020.

Toshihiko Eura
Senior Research Engineer, Cyber-World Laboratory, NTT Human Informatics Laboratories.

He received an M.E. from the Graduate School of Engineering at the University of Tokyo in 2005. He joined NTT EAST in 2005, where he was engaged in corporate user business development. From 2007 to 2019, he was engaged in network business development. From 2019 to 2021, he was engaged in the promotion project of the major international sporting event held in Tokyo in 2021.

Shinji Fukatsu
Senior Research Engineer, Supervisor, Cyber-World Laboratory, NTT Human Informatics Laboratories.

He received a Ph.D. in engineering from Osaka University in 2002 and joined NTT the same year. He has been engaged in R&D of human interfaces and video streaming services as well as in planning and development of video streaming services at NTT Plala and in promoting standardization and international development of information and communication technologies at the Ministry of Internal Affairs and Communications. He is currently engaged in R&D of remote-world infrastructure technology.