Acoustic XR Technology Merging Real and Virtual Sounds

Kenichi Noguchi, Hironobu Chiba, Tatsuya Kako, Shihori Kozuka, Yoshiaki Kurokawa, Yuki Watanabe, and Akira Nakayama

Abstract

With the spread of open-ear earphones that do not cover the ear, new listening experiences are being proposed that combines real ambient sounds with virtual sounds heard from earphones. At NTT, we call this merging of the real and virtual sounds through open-ear earphones "acoustic XR (extended reality) technology," which we are now developing with a view to actual services. In this article, we describe this technology with a focus on actual trials and touch upon future developments.

Keywords: XR, PSZ, open-ear earphones

1. Merging real and virtual sounds by open-ear earphones

NTT is researching and developing the Personalized Sound Zone (PSZ) as the ultimate private sound space and has developed design technology for an earphone that enables the hearing of sounds by only the user without the ear needing to be covered [1]. A variety of open-ear earphones that do not cover the ear have been appearing on the market and been spreading rapidly. At NTT, we have been focusing on a key feature of open-ear earphones, namely, the ability to naturally hear sounds from one's surroundings, and proposed and begun research and development on acoustic extended reality (XR) technology that merges virtual sounds heard from earphones and real sounds heard directly by the ear.

Acoustic XR technology extends the sounds that can be heard by adding sounds from earphones while listening to ambient sounds. For example, when attending a stage performance or concert, while the sounds generated by venue speakers are ordinarily heard at venue seats, there are still problems in reproducing sounds occurring near the audience, controlling the sense of sound direction and distance, and representing spatial sounds. In contrast, acoustic XR technology that merges sounds from loudspeakers and sounds from earphones will make it possible to present acoustics optimized for a variety of acoustic representations and for individual audience members, which has thus far been difficult to achieve. When attending a sports event at a stadium, acoustic XR technology will enable a user to enjoy commentary from earphones while experiencing the surrounding cheers of the crowd. Similarly, at a multilingual international conference, it will be possible to listen to the translated speech of other participants from earphones while simultaneously sensing nuances in their spoken speech.

As shown in **Fig. 1**, there are two key technical issues in acoustic XR technology: spatial sound for open-ear earphones and virtual-sound spatial rendering. Sounds generated with earphones generally create sound images localized inside the head. However, spatial sound for open-ear earphones can reproduce sounds generated from any position in space outside the head as if they were originating, for example, from the position of a real object. This makes all types of acoustic representations possible. This accommodates shifts in the wearing position of open-ear earphones or differences in individual ear shape and presents spatial sound through such earphones.

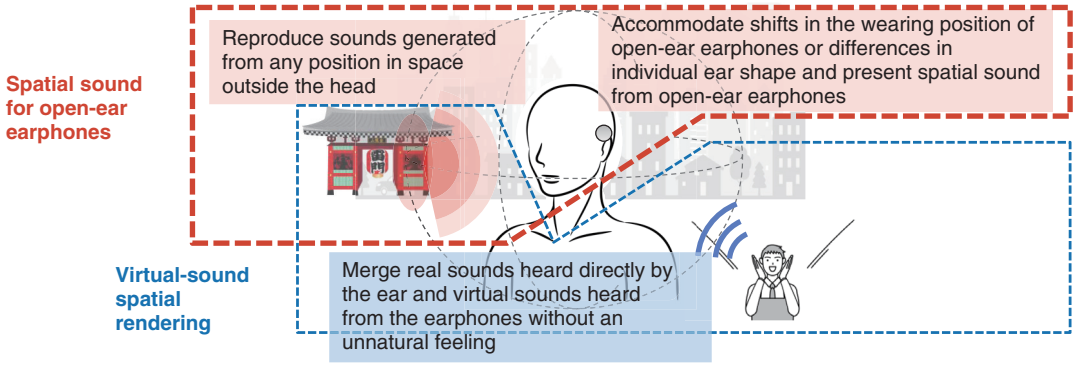


Fig. 1. Technical issues in acoustic XR technology.

Virtual-sound spatial rendering controls earphone-generated sound tailored to the user's cognitive characteristics and merges real sounds heard directly by the ear and virtual sounds heard from the earphones without an unnatural feeling.

NTT conducted several trials using acoustic XR technology in 2023, as described below.

2. Trial 1: Cho-Kabuki (Niconico Chokaigi 2023)

Cho-Kabuki is a performance that combines Kabuki, a traditional Japanese stage performance, with NTT advanced technology. "Cho-Kabuki Powered by NTT—Otogizōshi Koi No Sugatae—" was performed at the Niconico Chokaigi 2023 festival held at Makuhari Messe in Chiba Prefecture April 29–30, 2023. This program provided a spatial-sound performance in which real sounds in the venue and sounds flowing in the ear cross over each other. An audience member generally hears real sounds within the venue such as the actors' voices, music, and various sound effects emitted from venue speakers as well as shouts of admiration or encouragement from surrounding audience members (called *omuko*) directly at the ear. In this trial, we distributed open-ear earphones connected to a radio receiver to about 180 seats in front of the stage to provide those audience members with the experience of listening to sounds from earphones in addition to real venue sounds. These earphones played back sound effects synchronized with the program such as the hoofbeats of running horses, swishing sound of arrows flying through the air, and sound of wind. Combining these sounds with those from venue speakers can reproduce the sound of a horse running from left to right near the listener or the sound of an arrow flying above the listener, enabling a spatial sound production with a high sense of presence.

One problem with spatial sound using open-ear earphones is that the sound image may seem to be located upwards due to the acoustic characteristics of the enclosure, transmission characteristics of the sound propagating from the mounting position of the earphone to entrance of the external auditory canal, shape of the ear's surface, etc. To rectify this problem, we developed a compensating filter that cancels out this upward-sound-localization effect and applied it to sound effects that played back by the earphones (**Fig. 2**). This has enabled the generation of spatial sounds as intended by the content creator.

3. Trial 2: Audio guide for a satellite-transmission performance

A satellite-transmission performance was held to deliver high-quality sound of the 180th NTT EAST NHK Symphony Orchestra Concert held on November 2, 2023 at Tokyo Opera City Concert Hall (Shinjuku Ward, Tokyo) to Hokusai Hall in Obuse Town, Nagano Prefecture as a satellite venue. Hokusai Hall was equipped with a large screen for video playback and 5.1-channel sound equipment to play back high-quality sound. The performance also provided multi-angle delivery that enabled audience members to view the performance from any angle that they liked using a hand-held tablet or smartphone. Using open-ear earphones featuring minimal sound leakage based on NTT technology, we conducted a trial on delivering commentary on the music being performed. Since there was little sound leakage, it was possible to simultaneously hear the orchestra's performance and

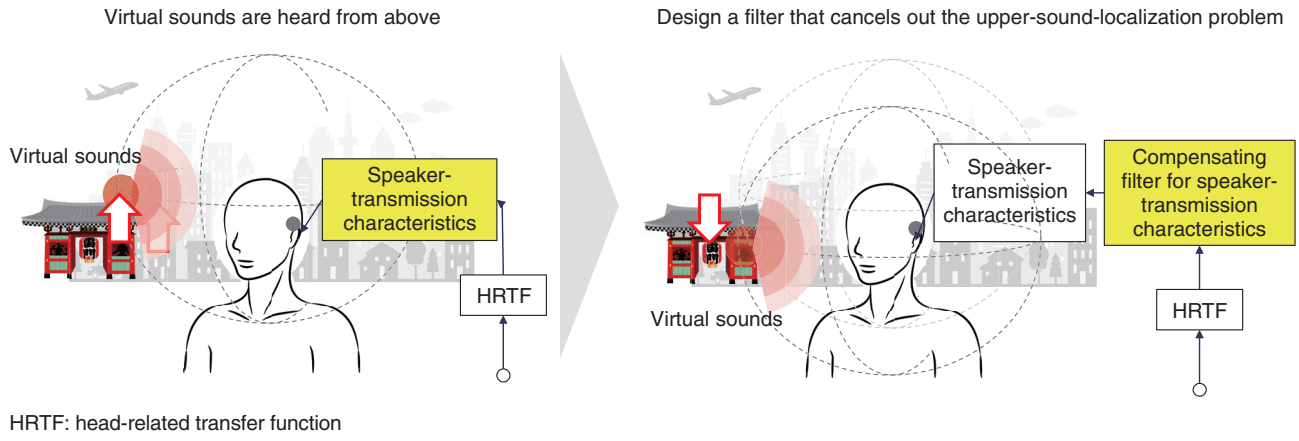


Fig. 2. Concept of applying a compensating filter to the localization position when wearing open-ear earphones.

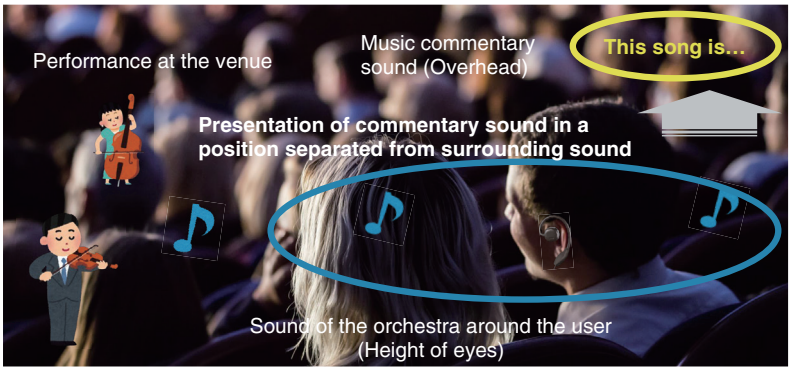


Fig. 3. Control the position of the sound image from the earphones.

commentary audio without bothering neighboring people and without needing to have the ear covered.

When using open-ear earphones, sounds are generally difficult to hear since real ambient sounds and sounds from the earphones overlap. To solve this problem, we controlled the position of the sound image from the earphones so that it appears to be coming from a position not covered by the real ambient sounds. This made it possible to present information by audio means even under conditions in which ambient sounds were being heard by the ear (**Fig. 3**). In this satellite-transmission performance, we used this technology to control the position from which commentary audio from the earphones could be heard somewhat above the user. This had the effect of separating the commentary audio from the sounds of the orchestra emitted from speakers, making it easier to hear and understand the commentary.

4. Trials 3: Audio guide for NTT History Center of Technologies

We conducted a trial of an audio guide using open-ear earphones at the NTT History Center of Technologies inside NTT Musashino R&D Center during the "NTT R&D FORUM 2023 — IOWN ACCELERATION" held November 14–17, 2023. We distributed to visitors a system that estimates user position by a smartphone sensor and plays back an audio guide automatically activated when approaching an exhibit from open-ear earphones connected to the smartphone. Although visitor guide services using earphones are provided at art galleries, museums, and other facilities, this trial stood out because of the following features.

(1) Appreciation of exhibits with open-ear earphones harmonized with surroundings

By not having the ear covered, ambient sounds can be heard naturally, enabling the user to understand surrounding conditions such as visitor congestion. Since earphone sound leakage is small, the user does not have to worry about bothering other visitors. The user can also naturally hear any sound emitted from building speakers or from the actual exhibits. Depending on the environment, it is also possible, for example, to enjoy exhibits while walking and chatting with a friend.

(2) Spatial sound heard as if coming from an exhibit

We enabled spatial sound that appears to be coming from an exhibit by making use of stereo sound playback and presenting signals that simulate the transmission characteristics of sounds propagating from the exhibit to the user's ears. We introduced technology that could compensate for the open-ear-earphone characteristic of localizing the sound image upwards. We also achieved spatial sound that appears to be coming from a stationary exhibit even while the user is moving by estimating user position in real time. To give a concrete example, a user may walk around the stationary exhibit of an old telegraph machine. The system would present sound effects mimicking the machine through the earphones by executing spatial sound processing in accordance with the user's position. This achieves acoustic direction that makes it appear as if sound is coming from the telegraph machine.

(3) Multilingual guide based on cross-lingual speech-synthesis technology

There is a growing demand for multilingual audio guides as the needs of inbound visitors to Japan increase. The cross-lingual speech-synthesis technology developed by NTT enables speech synthesis from speech data only in Japanese to a different language such as English or Chinese while maintaining the same voice quality. We prepared a Japanese/English audio guide using this cross-lingual speech-synthesis technology on the basis of Japanese speech data created by voice actors and enabled audio-guide switching on an app.

5. Future developments

We introduced acoustic XR technology for merging virtual sounds heard from earphones and real sounds heard directly by the ear and described recent trials. Open-ear earphones feature the ability to hear ambient sounds naturally, which suggests a variety of applications in scenarios other than tourism and entertainment such as business and everyday life. For example, they could be used to provide audio guidance to visually impaired persons. We actually tested the experience of walking around our office with a visually impaired person wearing open-ear earphones while listening to the audio guide. In interviews conducted after such experiences, we received similar comments to the following: "I was initially resistant to putting on earphones outside, but they hardly changed the way in which I heard outside sounds." Given that open-ear earphones do not cover the ear, they do not easily cause fatigue even after prolonged use. This should enable a variety of personalized acoustic XR services to be enjoyed depending on the user's current situation while wearing open-ear earphones all day. Going forward, we will work on solving whatever technical problems may arise in expected usage scenarios toward the provision of actual services.

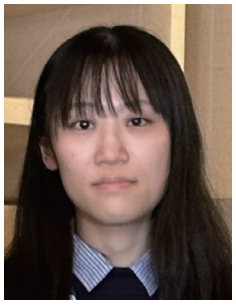
Reference

[1] NTT press release, "Developed earphone design technology that only the user can hear without blocking the ear—NTT developed a single speaker with sound wave control that delivers sound to the user while counteracting sound leakage to the surroundings," Nov. 9, 2022. https://group.ntt/en/newsrelease/2022/11/09/221109a.html

Kenichi Noguchi
Senior Research Engineer, Ultra-Reality Computing Group, NTT Computer and Data Science Laboratories.
He received a B.E. in electronic physics and M.E. in human system science from Tokyo Institute of Technology in 2001 and 2003 and joined NTT in 2003. His current research interests include audio-signal analysis and processing. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Acoustical Society of Japan (ASJ).

Hironobu Chiba
Research Engineer, Ultra-Reality Computing Group, NTT Computer and Data Science Laboratories.
He received a B.E. and M.E. in computer science from the University of Tsukuba, Ibaraki, in 2013 and 2015. He started his career as an R&D engineer at Pioneer Corporation in 2015 and joined NTT in 2019. His current research interests include hearable devices and acoustic-signal processing. He is a member of ASJ. He was the recipient of the Technical Development Award by ASJ in 2023.

Tatsuya Kako
Senior Research Engineer, Ultra-Reality Computing Group, NTT Computer and Data Science Laboratories.
He received a B.E. and M.E. in information science from Nagoya University, Aichi, in 2009 and 2011 and joined NTT in 2011. He has been engaged in research on microphone-array signal processing. He is a member of ASJ, IEICE, and the Institute of Electrical and Electronics Engineers (IEEE). He was the recipient of the Awaya Prize Young Researcher Award and the Technical Development Award by ASJ in 2023.



Shihori Kozuka
Research Engineer, Ultra-Reality Computing Group, NTT Computer and Data Science Laboratories.
She received a B.E. and M.E. in mathematical engineering and information physics from the University of Tokyo in 2020 and 2022 and joined NTT in 2022. Her current research interests include acoustic-signal processing and mathematical engineering. She is a member of ASJ.

Yoshiaki Kurokawa
Senior Research Engineer, Ultra-Reality Computing Group, NTT Computer and Data Science Laboratories.
He received a Ph.D. in applied chemistry from Waseda University, Tokyo, in 1997. He joined NTT the same year and has been engaged in research and development on magnetic and holographic recording. Since 2008 he has been engaged in research and development of communication systems and acoustic engineering. He is a member of IEEE and IEICE.

Yuki Watanabe
Research Engineer, Ultra-Reality Computing Group, NTT Computer and Data Science Laboratories.
She received a B.E. in electrical, information and physics engineering and M.E. in information science from Tohoku University, Miyagi, in 2021 and 2023 and joined NTT in 2023. Her current research interests include acoustic-signal processing and auditory psychology. She is a member of ASJ.

Akira Nakayama
Senior Research Engineer, Supervisor, Group Leader of Ultra-Reality Computing Group, NTT Computer and Data Science Laboratories.
He received an M.E. and Ph.D. in computer science from Nara Institute of Science and Technology in 1999 and 2007. After joining NTT in 1999, he has been engaged in robotics, computer-supported cooperative work, and recommendation and people-flow analysis. His current research interests are acoustics and signal processing. He is a member of the Information Processing Society of Japan, the Association for Computing Machinery, and the Robotics Society of Japan.