PSZ Active Noise Control and Desired Sound Selection Technologies for Creating a Comfortable and Safe Sound Environment in Vehicle Cabins

Noriyoshi Kamado, Tomoko Kawase, Masahiro Yasuda, Shoichiro Saito, Shihori Kozuka, Hiroaki Ito, and Akira Nakayama

Abstract

Technologies for muting unwanted sounds and for letting wanted (must-hear) sounds pass through, which are elements of a Personalized Sound Zone (PSZ), are described in this article. The elemental PSZ technology for confining sound, called spot-sound-reproduction technology, is first described. The combination of this technology with the new active noise control technology and sound event localization and detection technology—and examples of their applications—are then introduced. Applying these technologies to in-vehicle sound control makes it possible to provide drivers of advanced safety vehicles with "superior ears" that can see the blind spots not shown in mirrors, radar, and cameras. These PSZ technologies act as an "ear" that not only enhances driver and passenger comfort but also greatly enhances safety and reliability while providing a sophisticated connection among the driver, vehicle, vehicle exterior environment, and automotive society.

Keywords: active noise control (ANC), sound event localization and detection (SELD), advanced safety vehicle (ASV)

1. Two technologies necessary for controlling a sound space

Thanks to technological advances, the sound spaces surrounding us are evolving to become more convenient and comfortable. For a sound space, it is necessary to be able to block out the sounds one does not want to hear and transmit only the sounds one wants to hear, and manufacturers are introducing a variety of wearable products, such as earphones, to meet this need. However, these products have not been able to solve the following two major problems.

The first problem is the heavy strain on the ears when earphones are worn for long periods. Technically, blocking the ear is the most-effective (and cheapest) way to block out unwanted sounds, and most earphones with such a function are "in-ear" earphones, which—as the name suggests—are inserted into the ear canal. Long-term use of in-ear earphones can increase stress on the user due to pressure and increase the risk of ear-canal problems [1], and such issues raise concerns about the health of the ears, which are fundamental to human social activities. It can therefore be said that it is necessary to devise technology for controlling the sound space in a manner that blocks sounds the one does not want to hear without blocking one's ears.

The second problem is that in addition to the sounds one *wants* to hear, the sounds one *must* hear are not adequately considered. People can react and respond

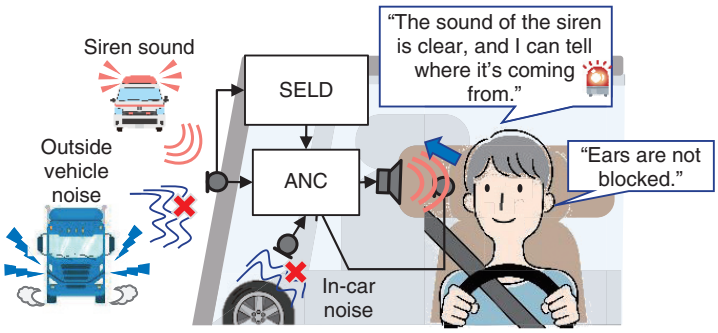


Fig. 1. Example applications of PSZ technology in vehicles.

to things that are happening in places invisible to the eye by hearing sounds from those places. For example, a person can avoid a bicycle approaching from behind by listening for the sound of the bicycle's ringing bell. These devices block out noise that does not need to be heard and must-hear sounds, and the person cannot always hear the sounds necessary for detecting danger. In other words, the ear loses its essential safety function of hearing sounds that must be heard. It can thus be said that the control of the sound space is required to ensure not only one hears the sounds one wants to hear but also one hears the sounds one must hear.

2. Two technologies for controlling the sound space in a vehicle: active noise control and sound event localization and detection

For vehicles, the above-described technical requirements are more pronounced than in other cases. Wearing a wearable device that blocks the ears is not only a possible violation of the Japan's Road Traffic Act but also a violation of certain other regulations. To create a comfortable sound space around a person traveling in a vehicle while ensuring their safety, it is thus necessary to be able to block the sounds one does not want to hear without blocking one's ears. The blocking of external sound with the vehicle's body and helmet, the increase in external noise including road noise while moving in a vehicle, and the fact that a vehicle is moving faster than a human make it difficult to hear sounds necessary to avoid danger, and that difficulty can be a cause of accidents. Consequently, the necessity to be able to hear the sounds that need to be heard is even greater, especially for advanced-safety vehicles (ASVs), which are essential for enhancing safety.

To meet the above-mentioned technological demands, we have been researching and developing elemental technologies for a Personalized Sound Zone (PSZ) that are highly integrated with spot-sound-reproduction technology and acoustic extended-reality (XR) technology. In this article, active noise control (ANC) technology, which suppresses noise without blocking the ears, and sound event localization and detection (SELD) technology are introduced (**Fig. 1**). SELD technology makes it easier to hear must-hear sounds to avoid danger even in environments where it is difficult to hear surrounding sounds.

3. ANC technology suppresses noise without blocking the ears

ANC technology for blocking sounds one does not want to hear without blocking the ears is explained. As mentioned above, popular earphones generally cover the ears to block sounds outside the ear. How commonly used in-ear earphones block sounds is shown in **Fig. 2(a)**. In-ear earphones are worn by inserting them into the ear canals. They act like earplugs by blocking the ear canals in a manner that makes it difficult to hear sounds outside the ear.

Sounds outside the ear, however, cannot be completely muted. Accordingly, as shown in **Fig. 2(b)**, two microphones (a reference microphone and error microphone) are fitted inside the earphone. On the basis of the sounds detected from moment to moment by these microphones, ANC reproduces sounds from the cancellation loudspeaker that eliminates the noise entering the ear. Therefore, sound-insulation performance improves. The reference microphone detects the ambient noise that is to be blocked, and the error microphone detects any sound missed by ANC

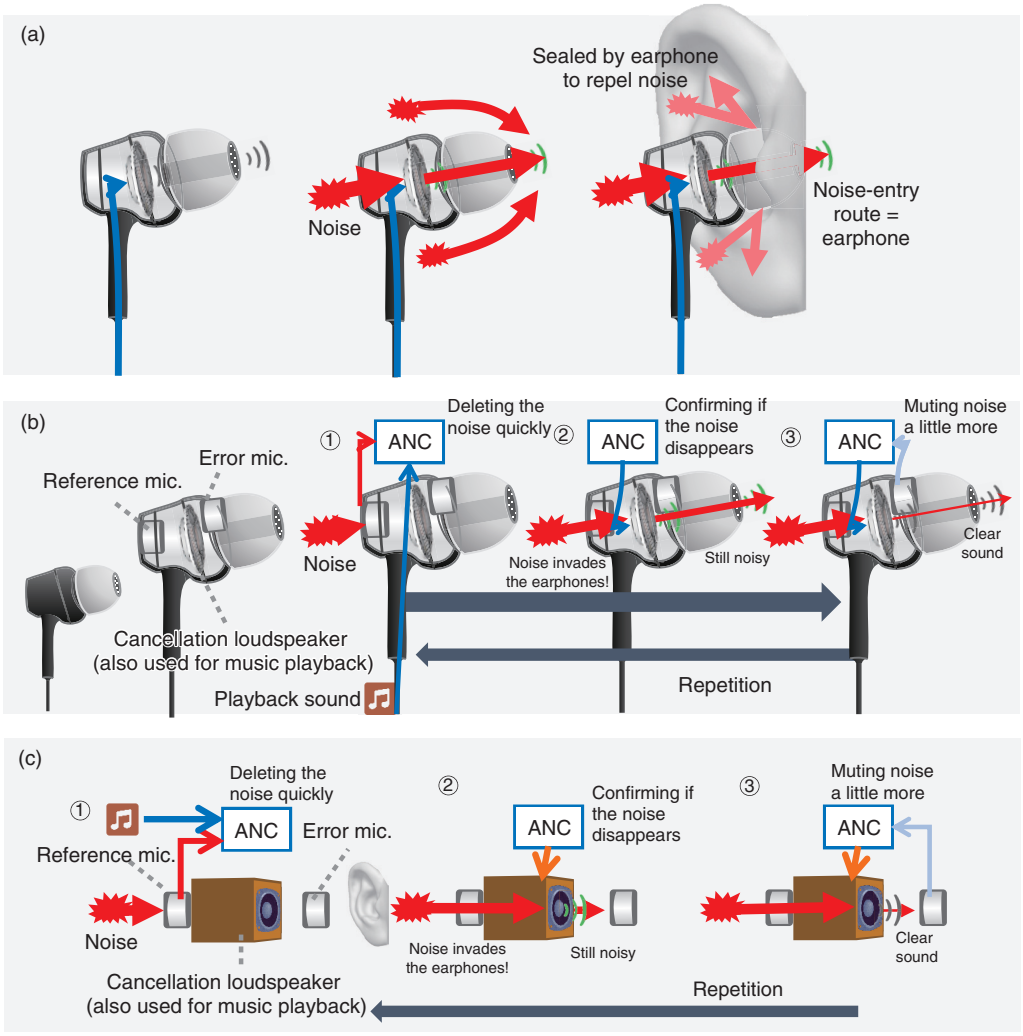


Fig. 2. (a) How in-ear earphones reduce noise. (b) How ANC works in in-ear earphones. (c) ANC that does not block the ear.

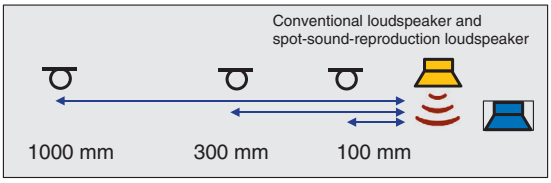
processing in the ear.

How to mute noise without covering the ears in the manner described above is considered. To avoid blocking the ear, the earphone loudspeaker must be positioned apart from the ear. However, a small loudspeaker, such as the one in an earphone, does not produce sufficient sound output, so a larger loudspeaker is required. The above-mentioned reference and error microphones should also be placed away from the ear because placing them near the ear, in the manner of in-ear earphones, would block the ear. Such a system is illustrated in **Fig. 2(c)**.

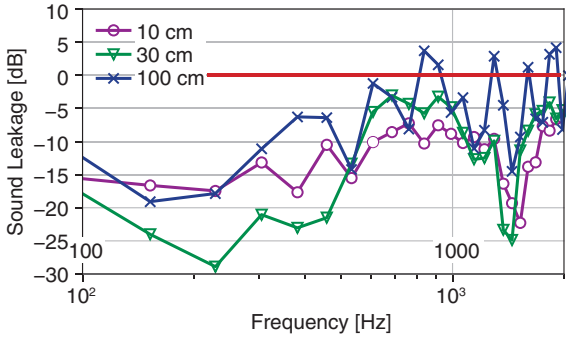
In consideration of the above circumstances, to suppress ambient noise without blocking the ear, it is necessary to cancel noise waves at the ear by using a loudspeaker and microphone placed apart from the ear. This configuration, however, faces the following three major problems:

(1) Control stability. The sound emitted by the loudspeaker (apart from the ear) to cancel out the noise is recorded by the reference microphone (which should record ambient noise only); as a result, so-called “howling” (also called “feedback”) occurs.

(2) As shown in Fig. 2(b), in-ear earphone ANC can correctly detect noise heard in the ear because the microphones are located near the ear entrance. On the contrary, as shown in Fig. 2(c), when the microphones are apart from the ear, they cannot correctly detect the noise



(a) System for measuring sound leakage from loudspeakers



(b) Sound-leakage comparison between conventional and spot-sound-reproduction loudspeakers

Fig. 3. Effectiveness of sound-leakage-suppression loudspeaker.

heard in the ear, and ANC outputs incorrect sound.

(3) Commercially available loudspeakers and computers take time to record and play back sound; thus, while the cancellation sound is being generated, the noise reaches the ears before being canceled. It is also undesirable to use large amounts of electricity in a vehicle's interior, so it is necessary to save power. Digital signal processors (DSPs) have conventionally been used for such applications. For ANC that does not block the ears, however, the problem of exceeding the computational power of the DSP must be solved. To solve problem (2), it is generally necessary to configure a large number of loudspeakers and microphones and implement signal processing for various compensations. This configuration requires a large amount of computation, and the cancellation sound cannot be generated in time.

To solve problem (1), it is necessary to reduce the sound leakage from the loudspeaker to the reference microphone. To meet that need, we developed a loudspeaker that applies the principle of the above-mentioned spot-sound-reproduction technology. The effect of reducing sound leakage is graphically shown in **Fig. 3**. This loudspeaker not only reduces the sound leakage to the entire surrounding of the loudspeaker but also creates an area where the sound leakage is very small, especially in the plane parallel to the diaphragm of the loudspeaker. The red line indicates the sound leakage of a conventional loudspeaker, and the other colored lines represent the amount of sound leakage from the loudspeaker using spot-sound-reproduction technology normalized by the sound pressure in front of the loudspeaker. Compared with the conventional loudspeaker, the spot-sound-reproduction loudspeaker suppresses sound leakage by several decibels to 30 decibels at 100 to 300 mm from the loudspeaker.

To solve problem (2), it is necessary to move the reference and error microphones closer to the ear. The spot-sound-reproduction loudspeaker reduces sound leakage, so by embedding it in the headrest of a vehicle, the reference microphone can be moved closer to the ear. The reference microphone is then able to detect sounds similar to noise heard with the ear.

Unlike the reference microphone, the error microphone does not suffer from the howling problem, so

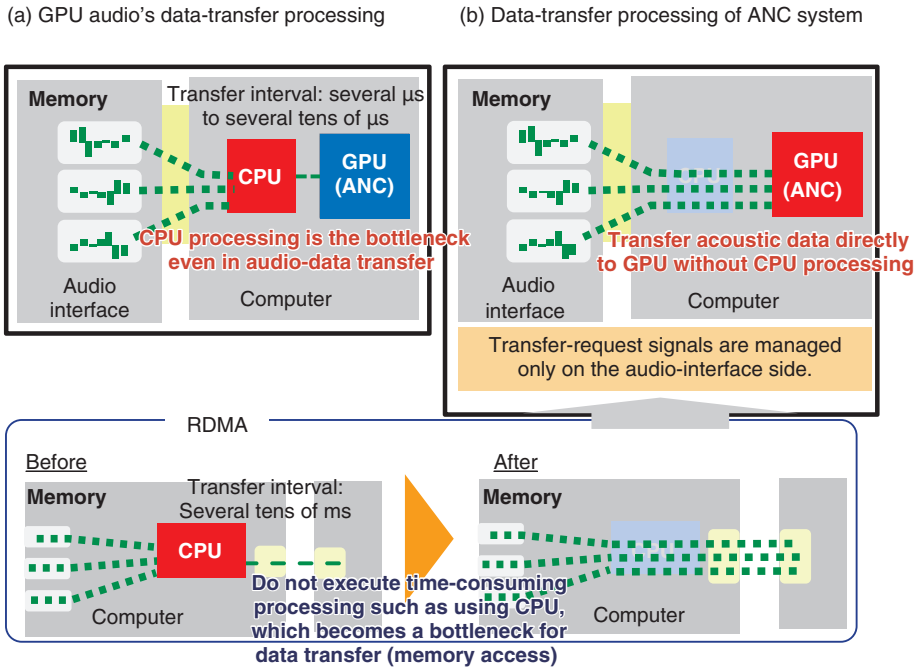


Fig. 4. GPU direct audio method for transferring data at high speed and low latency in short intervals with low power consumption.

it can be placed closer to the ear than the reference microphone. However, it is impossible to place the error microphone near the ear without blocking the ear. Therefore, to estimate the noise at the ear with the microphone slightly away from the ear, signal processing is required. This processing typically requires the placement of multiple error microphones, and it must be completed within a very short period (a few hundred microseconds) after the noise is detected by the reference microphone and before it reaches the ear.

For the above reasons, power-saving hardware that executes signal processing at high speed and with ultra-low latency is essential for ANC that does not block the ears. We solved this problem by applying general-purpose computing on graphics processing units (GPGPUs), which are widely used in the network and video-processing fields, to acoustic processing and optimizing it.

The principle of using GPGPUs for ANC is illustrated in **Fig. 4**. GPGPUs are known for their high speed, low latency, and low power consumption per unit of processing. However, as shown in Fig. 4(a), data cannot be directly input to and output from audio devices, and a central processing unit (CPU) must act as an intermediary between them, increasing processing delay. Therefore, by using remote direct memory access (RDMA) technology, as shown in Fig. 4(b), the audio signals of the microphone and loudspeaker are connected directly to the GPGPU without going through a CPU. The method was optimized to achieve processing of large amounts of small data-transmission packets (frame bursts) not handled in other signals and a distinct real-time nature unique to acoustic processing, which cannot tolerate even a 1-µs delay.

The result of the above-described RDMA application is hardware that can transfer acoustic data in about 1/50th the time required by conventional hardware and process large amounts of acoustic signals in real time and with low power consumption. This hardware not only enables ANC that does not block the ears, which has been difficult to achieve with DSP, but also enables the introduction of deep-learning technology, which is said to be difficult to implement in ANC due to its large computational load and significant processing delays.

An example of a test vehicle equipped with these technologies is shown in **Fig. 5**. In the test vehicle, loudspeakers with the spot-sound reproduction function are installed on both sides of the headrests of all seats, in positions that do not obstruct the driver's/passenger's line of sight, and reference and error

Fig. 5. Example of implementing an ANC system that does not block the ears.

microphones are placed around the loudspeakers. The shape of the headrest is designed to prevent the spot-sound-reproduction capability from deteriorating due to the spot-sound-reproduction loudspeakers being mounted inside the seat's headrests, thereby enhancing the accuracy of noise suppression near the ear. This configuration improves the comfort of the people inside the vehicle interior.

4. SELD technology enables one to hear only the sounds one must hear

As mentioned above, ANC technology can suppress unpleasant sounds when the vehicle is running; however, all sounds detected with the error microphones are subjected to suppression, which can lead to accidents by making it difficult to hear the sounds that must be heard to avoid danger. To make it possible to hear the sounds one needs to hear, it is therefore necessary to develop new sound-transmission technology. To meet this need, we focused on SELD technology.

SELD is overviewed in **Fig. 6**. SELD technology estimates when, where, and what happened from sound signals observed with microphones. By using this technology, it is possible to detect the sound (including its direction of arrival (DOA)) that the driver truly needs from the various sounds input into the microphones. SELD technology is now generally based on end-to-end deep-learning technology, which internally estimates the DOA corresponding to "where" and sound event detection (SED) corresponding to "what."

Since SELD technology uses deep learning, it requires a large amount of data related to necessary sounds that must be heard for its training. For example, an application of SELD technology—in which a siren is sounding in the driver's blind spot—is shown in Fig. 6. In this situation, the siren is quickly detected, and the driver is notified of the direction of the siren so that they can take the appropriate evasive action. Ideally, sirens from all directions and in all blind spots should be recorded and used for training the deep-learning model while considering all locations and situations (surrounding vehicles, buildings, weather conditions, etc.). However, it is difficult to record such a large volume of sound data comprehensively.

People can infer, to some extent, the direction of sound arrival, even in the presence of environmental differences and changes in sound due to self-motion [4–6]. In other words, a person can select the information necessary to estimate the direction of sound arrival from the information contained in the sound. Given this fact, we considered enabling SELD technology to imitate this human ability [7–9].

We developed echo-aware feature-refinement (EAR) - comprehensive anechoic data and sparse multi-environment data (CASM) technology [2] and motion-aware feature-refinement (MAR) technology [3] to enable SELD technology to imitate such human abilities. These technologies enable SELD technology

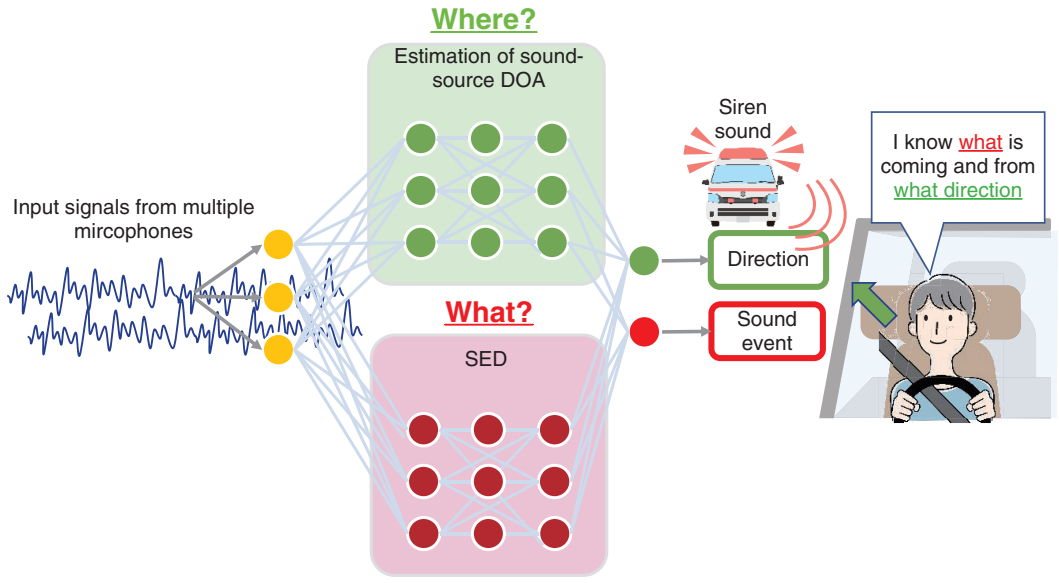


Fig. 6. Overview of SELD technology.

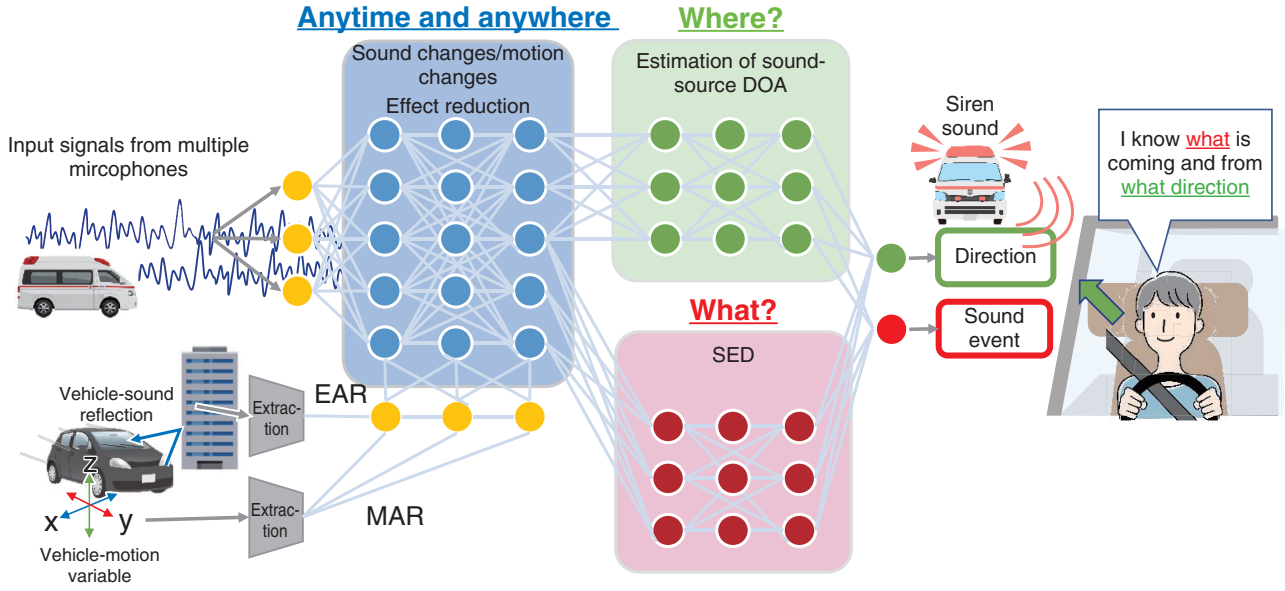


Fig. 7. Overview of automotive SELD technology applying EAR-CASM and MAR.

to operate robustly even when the environment changes or the user moves. It has thus become possible to apply SELD technology to mobile environments, such as cars, that were previously unrealistic at reasonable cost.

An overview of SELD technology for automotive applications applying EAR-CASM and MAR technologies is given in **Fig. 7**. The EAR-CASM technology provides the neural network with echoes of sounds emitted from the SELD-equipped vehicle (e.g., the car's running sound and sonar sound from its sensors) that are reflected to its surrounding as a cue for learning sounds not included in SELD technology's training data. The "improved" SELD

technology works like the human ear, i.e., it suppresses the effects of unknown environmental sounds from these echoes by using the sounds it has learned up to that time. Moreover, MAR reduces the effect of changes in sound related to the vehicle's own motion by providing the neural network with various inputs from sensors (such as acceleration sensors fitted in the vehicle) and information cameras as information about the vehicle's motion.

Regarding EAR-CASM technology, exhaustive recordings of sounds from various environments are no longer necessary, and SELD technology can be implemented at a realistic cost. By applying EAR-CASM and MAR technologies in conjunction with ANC technology to ASVs, we have been creating a sound space for safe and comfortable transportation in which sounds one does not want to hear can be blocked—without blocking one's ears, and sounds one wants to hear—as well as sounds one must hear—can be heard.

References

[1] C. Mukhopadhyay, S. Basak, S. Gupta, K. Chawla, and I. Bairy, "A Comparative Analysis of Bacterial Growth with Earphone Use," OJHAS, Vol. 7, No. 2, April 2008.

[2] M. Yasuda, Y. Ohishi, and S. Saito, "Echo-aware Adaptation of Sound Event Localization and Detection in Unknown Environments," Proc. of the 47th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022), pp. 226–230, Singapore, May 2022. https://doi.org/10.1109/ICASSP43922.2022.9747603

[3] M. Yasuda, S. Saito, A. Nakayama, and N. Harada, "6DoF SELD: Sound Event Localization and Detection Using Microphones and Motion Tracking Sensors on Self-motioning Human," Proc. of the 49th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024), Seoul, Korea, Apr. 2024.

[4] R. Gao, C. Chen, Z. Al-Halah, C. Schissler, and K. Grauman, "VisualEchoes: Spatial Image Representation Learning through Echolocation," Proc. of the 16th European Conference on Computer Vision (ECCV 2020), Aug. 2020. https://doi.org/10.1007/978-3-030-58545-7_38

[5] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of Room Geometry from Acoustic Impulse Responses," IEEE Trans. on Audio, Speech, and Lang. Process., Vol. 20, No. 10, pp. 2683–2695, 2012. https://doi.org/10.1109/TASL.2012.2210877

[6] L. D. Rosenblum, M. S. Gordon, and L. Jarquin, "Echolocating Distance by Moving and Stationary Listeners," Ecological Psychology, Vol. 12, No. 3, pp. 181–206, 2000. https://doi.org/10.1207/S15326969ECO1203_1

[7] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-related Transfer Functions on the Spatial Perception of a Virtual Speech Source," J. Audio Eng. Soc., Vol. 49, No. 10, pp. 904–916, 2001.

[8] Y. Iwaya, Y. Suzuki, and D. Kimura, "Effects of Head Movement on Front-back Error in Sound Localization," Acoust. Sci. & Tech., Vol. 24, No. 5, pp. 322–324, 2003. https://doi.org/10.1250/ast.24.322

[9] B. C. J. Moore, "An Introduction to the Psychology of Hearing (3rd ed.)," Academic Press, 1989.

Noriyoshi Kamado
Senior Research Engineer, Ultra-Reality Computing Group, NTT Computer and Data Science Laboratories.

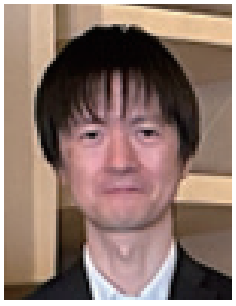
He received an M.E. in electrical and electronic systems engineering from Nagaoka University of Technology, Niigata, in 2009 and Ph.D. from Nara Institute of Science and Technology in 2012. He joined NTT in 2012 and NTT DOCOMO in 2015, where he has been working on speech-signal processing technologies for a speech-recognition system. He is a member of the Acoustical Society of Japan (ASJ), the Institute of Electrical and Electronics Engineers (IEEE), and the Audio Engineering Society (AES).

Tomoko Kawase
Research Engineer, Ultra-Reality Computing Group, NTT Computer and Data Science Laboratories.

She received a B.E. and M.E. in system design engineering from Keio University, Kanagawa, in 2011 and 2013 and Ph.D. in information engineering from Tsukuba University in 2018. Since joining NTT in 2013, her research interests have included microphone array signal processing and speech enhancement. She is a senior member of IEEE and a member of ASJ.

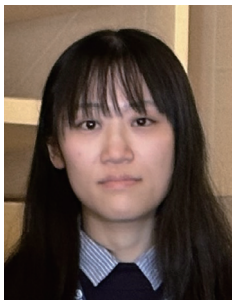
Masahiro Yasuda
Research Engineer, Ultra-Reality Computing Group, NTT Computer and Data Science Laboratories.

He received a B.E. and M.E. in physics from Tokyo Institute of Technology in 2017 and 2019. He joined NTT in 2019, and his current research interests include acoustic-signal processing and machine learning. He is a member of IEEE and ASJ.



Shoichiro Saito
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 the University of Tokyo in 2005 and 2007. Since joining NTT in 2007, he has been involved with the research and development of acoustic signal processing systems, including acoustic echo cancellers, hands-free telecommunication, and ADS. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and ASJ.



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.

Hiroaki Ito
Senior Research Engineer, Ultra-Reality Computing Group, NTT Computer and Data Science Laboratories.

He received a B.E. in electronics and M.E. in information science from Nagoya University, Aichi, in 2007 and 2009 and joined NTT in 2009. His current research interests include acoustic-signal processing and sound-field control. He is a member of ASJ, IEICE, the Institute of Image Information and Television Engineers, and AES.

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.