Media-processing Technologies for Ultimate Private Sound Space

Masahiro Fukui, Shoichiro Saito, and Kazunori Kobayashi

Abstract

NTT Media Intelligence Laboratories has been researching and developing media-processing technologies to create ultimate private spaces for promoting digital transformation in diverse spaces, including work-style reform. This article focuses on sound, which is one of the most important elements in achieving private space and introduces our aims in establishing the following three technologies: technology to understand surroundings from sound (event-detection and scene-identification techniques), active-sound-control technology to make the sound heard only by listeners who want to hear, and active-noise-control technology to eliminate undesired sounds. This article describes our, the authors', efforts regarding these technologies.

Keywords: event detection, active sound control, active noise control

1. Introduction

Due to the work-style reforms promoted by the Japanese government and the impact of the novel coronavirus, the traditional way of working that requires employees to come to the office has been reconsidered, and flexible work styles that are not bound by location or time have been attracting attention. An important part of this new work style is the ability to create a comfortable working environment, no matter where one is.

Let us consider telecommuting (remote working). There are various types of noise in the home, such as from the air conditioner, traffic outside, and sometimes chimes to announce deliveries. If there are family members present, there may be their voices and sounds from the television (TV). Such noise includes sounds that telecommuters do not want to hear. However, in some situations, chimes and a baby crying may be sounds one wants to hear. When a conference call is held at home, we do not want the noise generated near us to reach the other party. We also do not want those near us to hear the voice from the other party. If we can create an ultimate private sound space in which telecommuters can hear only what they want to hear, e.g., voices from their telecommunication partners, they will be able to work comfortably at home.

The goal of NTT Media Intelligence Laboratories is to create the ultimate private sound space, which we call a Personalized Sound Zone (PSZ), by accurately collecting sound information and understanding the situation. NTT Media Intelligence Laboratories has accumulated a large amount of knowledge on sound-collection technology and is currently working on further developing sound situational understanding and control technologies. The following describes the three major technologies (**Fig. 1**).

2. Technology to understand surroundings from sound (event-detection and scene-identification techniques)

People want to hear different sounds depending on the situation. For example, a user may want to hear his/her dog bark when at home, but may not want to hear the barking of other dogs when away from home. In such a case, if we can detect the sound of a dog barking while the user is away from home, we can judge that the sound as undesired.

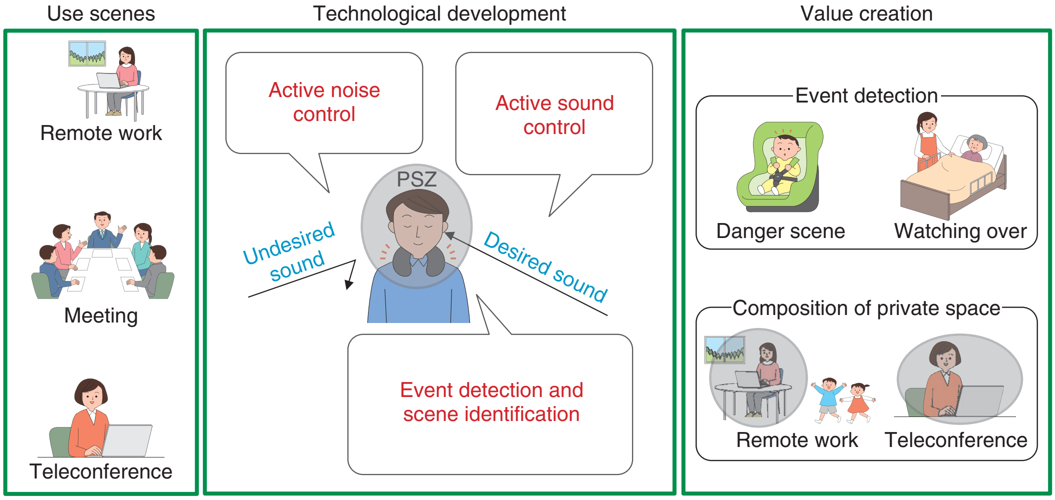

Fig. 1. Concept of Personalized Sound Zone (PSZ).

Thus, to achieve a PSZ, it is important to selectively convey sounds and situations to the user according to the situation, rather than simply suppressing all ambient sounds. To do so, it is necessary to be aware of the environment surrounding the user. For this purpose, we are developing an event-detection technique for simultaneous estimation of information, such as "when," "what," and "where," and a scene-identification technique for estimating the meaning of information such as "in what situation" and "why."

One of the difficulties with event detection is that the sound changes in various ways depending on the surrounding environment, even if the sound reaches the same location. For example, deep neural networks (DNNs) have recently become the mainstream method for determining where sounds are generated, but due to the diversity of the environment, even DNNs cannot be 100% effective. We are working on improving estimation accuracy by using the spatial symmetry of the sound field and combining it with physical-quantity estimation methods [1–3]. We are also investigating a method for satisfying the requirement of detecting only specific events with fast computation and low complexity [4].

The goal with the scene-identification technique is to estimate the information about the user's situation as higher-level information than events and sound-source locations. For example, this technique will estimate not only the sound of a car running but also the user's situation; therefore, it can judge whether the sound should be suppressed or alert the user according to his/her situation. We are currently developing a sound-description-generation technique [5] for describing sound signals in natural language as an elemental technology of the scene-identification technique.

3. Active-sound-control technology to make the sound heard only by listeners who want to hear

Listeners have been using earphones and headphones to listen to sound without affecting their surroundings. However, there are many problems, such as the inconvenience of wearing/carrying a device, possibility of fatigue and hearing loss due to prolonged use, and difficulty in detecting the situation and possible danger around when wearing/carrying such a device. Therefore, if localized sound-zone generation is possible so that only the target listener can hear without using earphones or headphones, these problems can be solved (**Fig. 2**). NTT Media Intelligence Laboratories is engaged in research and development (R&D) of both software and hardware to develop active-sound-control technology for generating a localized sound zone. The following sections discuss issues and approaches regarding this technology.

3.1 Efforts to improve software performance

Although localized sound-zone generation requires multiple loudspeakers, its control is implemented using software and uses signal processing called active sound control. This technology requires more

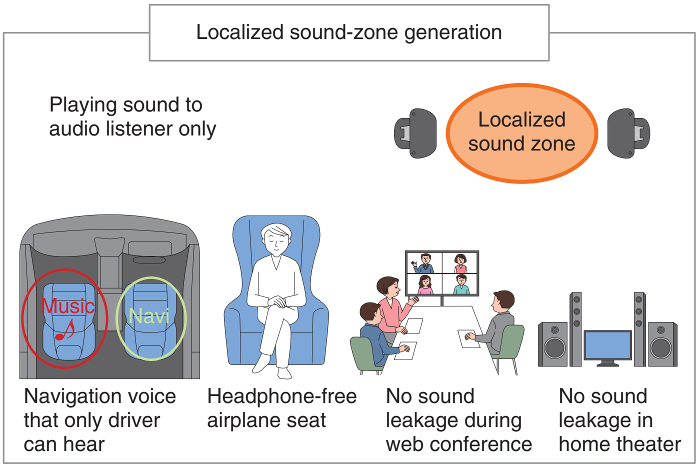
Localized sound-zone generation
Playing sound to audio listener only
Localized sound zone
Music
Navi
Navigation voice that only driver can hear
Headphone-free airplane seat
No sound leakage during web conference
No sound leakage in home theater

Fig. 2. Application range of active-sound-control technology.

loudspeakers when the upper frequency range is set higher. Also, there may be restrictions on the placement of each speaker. PSZs are designed for individual users, and the number and degree of freedom of loudspeaker placement are severely restricted, unlike general problem settings. For example, at home, the space where the speakers can be installed is limited, such as around a computer desk. The aim with PSZs is to achieve localized sound-zone generation with a small number of loudspeakers and limited space. Active-sound-control technology identifies all the conditions necessary for filter design and optimizes the overall filter to satisfy all conditions simultaneously.

3.2 Efforts to improve hardware performance

In addition to researching signal processing, we are also investigating the loudspeaker arrangement that minimizes sound leakage with a limited number of loudspeakers and installation locations as well as hardware that has a greater attenuation of sound level as it becomes further away from the loudspeakers. We are also engaged in R&D of small speakers for bass reproduction. Bass is important to maintain high sound quality. The physical size of the loudspeaker body is necessary to reproduce bass at a sufficient volume. However, as described above, it is not practical to use a large loudspeaker because of the space constraint of PSZs. We aim to improve the hardware so that the bass limit of small speakers can be lowered more than ever before.

4. Active-noise-control technology to eliminate undesired sounds

A PSZ uses the event-detection and scene-identification techniques to discriminate between incoming sounds to create a space where undesired sounds are not heard. Currently, noise cancellation in earphones and other widely used devices is easy to achieve because the space to muffle the sound is small and fixed. However, wearing earphones for a long period causes discomfort, such as earache. If we can develop technology to muffle undesired sounds using a device that does not need to be worn, it will be more convenient and expand the range of applications (**Fig. 3**).

Active-noise-control technology that muffles sound in a space consists of a control loudspeaker that generates the control sound, error microphone that detects the error signal at the control point, reference microphone that references the noise signal, and controller that uses an adaptive algorithm to produce the control sound. If the error signal detected with the error microphone is small, undesired sounds are reduced.

The more loudspeakers there are that produce controlled sounds, the more points can be controlled and the easier it is for active noise control to eliminate undesired sounds. However, when considering use in the home, a small number of speakers is desirable. Considering the housing conditions in Japan, these loudspeakers and microphones will be located close to each other. Therefore, problems may occur, such as degradation in performance due to the control sound going around the reference microphones, which was

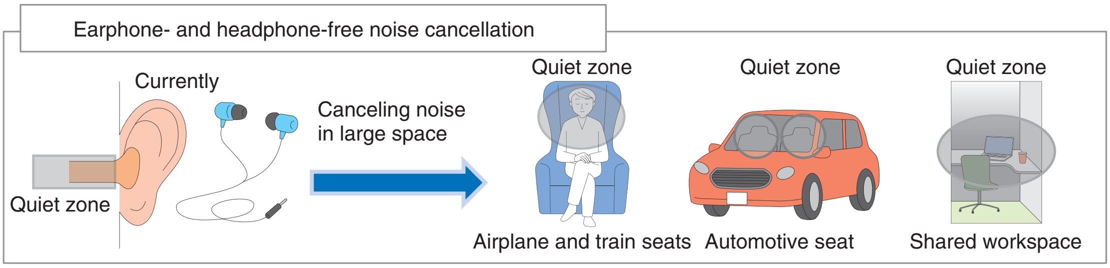

Fig. 3. Application area of active-noise-control technology.

not envisioned in the previous active-noise-control technology.

Undesired sound depends on the individual and environment. For example, people prefer not to hear the sound of a car when they are indoors, but a pedestrian feels safer if he/she can hear car noise. After detecting the sound with the event-detection and scene-identification techniques, technology is also needed to determine the necessity of sound according to the situation, for example, whether the sound is necessary at the moment. To achieve a space where the listener can hear only what the listener wants to hear, multiple technologies are needed, and they must be coordinated at a high level.

5. Future prospects

This article outlined PSZs and described the current status of its elemental technologies, such as technology to understand surroundings from sound (event-detection and scene-identification techniques), active-sound-control technology, and active-noise-control technology. There are still many technical issues to be addressed, and NTT Media Intelligence Laboratories will continue to conduct R&D on this topic. We will also collaborate with other organizations both inside and outside the group to achieve PSZs.

References

[1] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, “First Order Ambisonics Domain Spatial Augmentation for DNN-based Direction of Arrival Estimation,” Proc. of the 4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2019), New York, USA, Oct. 2019.

[2] M. Yasuda, Y. Koizumi, S. Saito, H. Uematsu, and K. Imoto, “Sound Event Localization Based on Sound Intensity Vector Refined by DNN-based Denoising and Source Separation,” Proc. of the 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020), May 2020.

[3] R. Sato, K. Niwa, and K. Kobayashi, “Ambisonic Domain DNN Preserving Physical Symmetry and Its Application to Sound Event Detection and Direction of Arrival Estimation,” The 2020 Autumn meeting of the Acoustical Society of Japan, Sept. 2020 (in Japanese).

[4] S. Murata, S. Saito, K. Kobayashi, and A. Nakagawa, “A Study of Lightweight Sound Event Detection Method Based on Decision,” The 2020 Autumn meeting of the Acoustical Society of Japan, Sept. 2020 (in Japanese).

[5] Y. Koizumi, R. Masumura, K. Nishida, M. Yasuda, and S. Saito, “A Transformer-based Audio Captioning Model with Keyword Estimation,” Proc. of INTERSPEECH 2020, Oct. 2020.

Masahiro Fukui

Senior Research Engineer, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.

He received a B.E., M.E., and Ph.D. in information science from Ritsumeikan University, Shiga, in 2002, Nara Institute of Science and Technology in 2004, and Ritsumeikan University in 2018. Since joining NTT in 2004, he has been engaged in research on acoustic echo cancellers and speech coding. He received the best paper award of the ICCE conference and the technical development award from the Acoustic Society of Japan (ASJ) in 2014. He is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the ASJ.

Shoichiro Saito

Senior Research Engineer, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.

He received a B.E. and M.E. from the University of Tokyo in 2005 and 2007. Since joining NTT in 2007, he has been engaging in research and development of acoustic signal processing systems, including acoustic echo cancellers, hands-free telecommunication, and anomaly detection in sound. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), IEEE, and ASJ.

Kazunori Kobayashi

Senior Research Engineer, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.

He received a B.E., M.E., and Ph.D. in electrical and electronic system engineering from Nagaoka University of Technology, Niigata, in 1997, 1999, and 2003. Since joining NTT in 1999, he has been engaged in research on microphone arrays, acoustic echo cancellers, and hands-free systems. He is a member of IEICE and ASJ.