Pursuing Both Basic Research and International-standardization Activities to Make Technology Useful to Society

Noboru Harada
Senior Distinguished Researcher,
NTT Communication Science Laboratories

Abstract

As remote conferencing is becoming commonplace and the term “metaverse” is frequently appearing in the media, society’s attention is increasingly focused on communication using video and audio. Although the role of video in such communication tends to gain greater interest, the role of audio—for example, in creating a sense of immersion—is also significant. Noboru Harada, a senior distinguished researcher at NTT Communication Science Laboratories, has been researching speech and acoustic communication for more than 20 years and is currently focusing on research and standardization of natural and highly functional speech and acoustic communication technology to achieve, for example, selective transmission of sounds you want to convey and sounds you want to hear. We interviewed him about his research accomplishments, standardization activities, and attitude as a researcher.

Keywords: speech/audio coding, immersive communication, semantic segmentation

Contributing to society through basic research and international standardization of speech and acoustic communication

—Would you tell us about the research you are currently conducting?

In 2022, I set a new research theme “research and standardization of natural and highly functional speech/acoustic communication technology,” and I’m researching technologies to (i) achieve highly functional immersive communication that supports multipoint conferences that are natural to people and selectively conveys only the sounds people want to hear or convey to others and (ii) create a speech and acoustic communication environment in which the quality of experience (QoE) is naturally and automatically improved according to individuals and situations.

When people are having a face-to-face conversation, even in the presence of ambient noise, they selectively hear the voice of the person they are

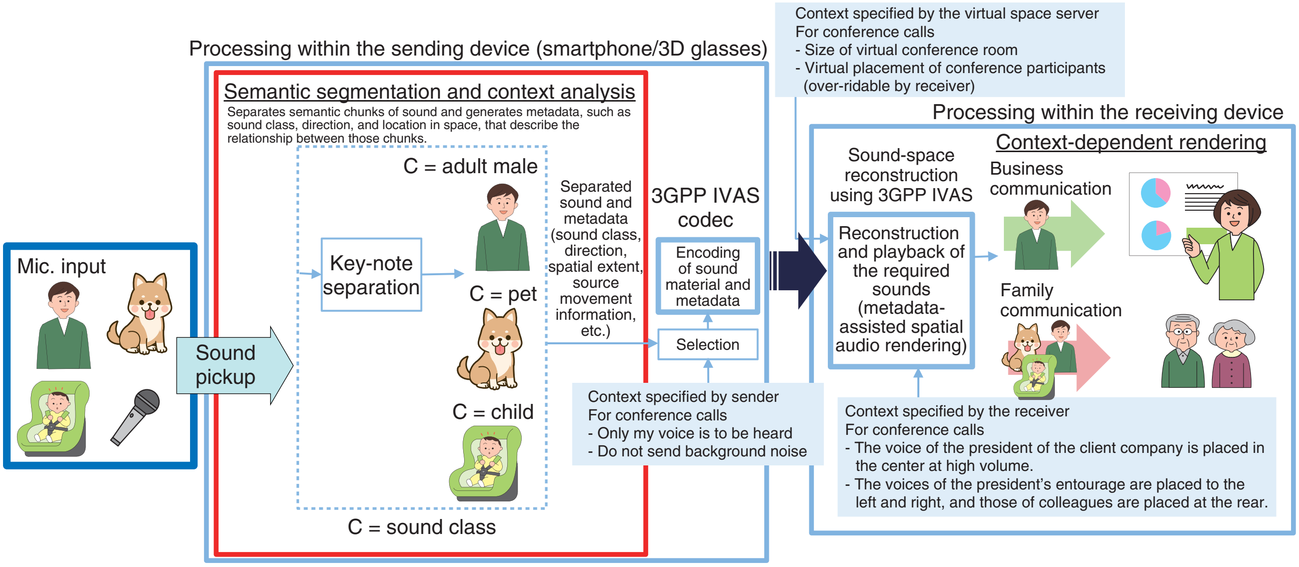

Context specified by the virtual space server
For conference calls
- Size of virtual conference room
- Virtual placement of conference participants (over-ridable by receiver)
Processing within the sending device (smartphone/3D glasses)
Semantic segmentation and context analysis
Separates semantic chunks of sound and generates metadata, such as sound class, direction, and location in space, that describe the relationship between those chunks.
Mic. input
Sound pickup
Key-note separation
C = adult male
C = pet
C = child
C = sound class
Separated sound and metadata (sound class, direction, spatial extent, source movement information, etc.)
3GPP IVAS codec
Encoding of sound material and metadata
Selection
Context specified by sender
For conference calls
- Only my voice is to be heard
- Do not send background noise
Processing within the receiving device
Sound-space reconstruction using 3GPP IVAS
Reconstruction and playback of the required sounds (metadata-assisted spatial audio rendering)
Context-dependent rendering
Business communication
Family communication
Context specified by the receiver
For conference calls
- The voice of the president of the client company is placed in the center at high volume.
- The voices of the president's entourage are placed to the left and right, and those of colleagues are placed at the rear.

Fig. 1. Example applications of semantic-segmentation technology to immersive communication.

talking to. When people are participating in remote conferences at home, however, the microphone faithfully picks up ambient sounds that the speakers do not want to be heard, and those sounds are played through the loudspeakers of the other participants. Conversely, when family members (such as parents and grandchildren) who live far apart are remotely communicating, these ambient sounds enhance the intimacy of the communication. In other words, in communication, we want to convey and hear certain sounds.

Certain communication methods are used for remote environments, and others are used for "real" (i.e., face-to-face) environments. The use of remote environments was accelerated by the COVID-19 pandemic, but with the end of the pandemic, many situations that mix remote and real environments have emerged. Therefore, developing communication methods for such a hybrid environment is an issue.

We can address the above issues by selectively communicating only the sounds we want to convey or hear in accordance with the purpose and situation. To achieve such highly functional immersive communication, it is necessary to understand the semantic structure of input acoustic signals and extract acoustic objects (i.e., semantic chunks of sound, including metadata) by applying technologies for understanding the sound environment such as acoustic-event detection and sound-source-direction estimation. The key technology to meet this requirement is to automatically acquire representations for executing such extraction from data.

Reconstructing the extracted acoustic objects and applying them to, for example, transmission and virtual-reality (VR) and augmented-reality (AR) playback is being investigated. For VR/AR communication, "semantic segmentation" is an important technology to separate acoustic objects from multiple mixed sound inputs and output them with class and spatial information (**Fig. 1**). Since some use cases do not necessarily require high quality or the extraction of all acoustic objects, highly functional immersive communication is implemented by (i) estimating QoE in accordance with the use case (by modeling the sound-quality-evaluation criterion on the basis of human-auditory perception) and (ii) automatically improving and tuning the sound quality and acoustic objects to be extracted to meet the criterion.

I believe that these technologies for highly functional immersive communication can be used for not only conference calls but also remote medical care, monitoring the elderly, remote control of heavy machinery, and other use cases where sound is heard remotely. With these use cases in mind, I'm aiming for practical application of these technologies through international standardization and other means.

—Before setting this new theme, what type of research have you conducted thus far?

Since joining NTT in 1997, with the exception of a period of development work at an operating company and research management, I have worked on speech/acoustic signal processing, coding, and its international standardization. Recently, I have specialized in self-supervised learning and other representation learning techniques and in applying these techniques to understand sound environments such as acoustic-event detection and anomalous-sound detection.

An achievement representing that work includes the technical proposal and standardization of MPEG-4 Audio Lossless Coding (ALS)*, a lossless coding method for acoustic signals. MPEG-4 ALS is implemented in encoder equipment manufactured by NTT Electronics and used for high-resolution music distribution and other applications. In 2018, in cooperation with NTT DATA and other operating companies, we practically applied our technology to detect anomalies from sounds made by machines and other equipment. This technology is being used to (i) monitor vehicle abnormalities by analyzing train-running sounds recorded on the ground and (ii) identify and detect damage and deterioration of equipment in facilities such as power plants.

I have contributed to international standardization through the following activities. At MPEG in ISO/IEC, I helped standardize the MPEG archive format in audio coding as a chair of the MPEG-Audio Ad Hoc Group and helped standardize MPEG-4 ALS as an editor. At IEC, I helped establish the IEC61937-10 transmission standard as a project leader and editor, and at the International Telecommunications Union - Telecommunication Standardization Sector (ITU-T), I served as an editor on the audio coding standard, ITU-T Recommendation G.711.0. At the Internet Engineering Task Force, I served as an editor on RFC 7655, which specifies payload formats in the Real-time Transport Protocol.

At the 3rd Generation Partnership Project (3GPP), an international standardization body for mobile communications, I also contributed to the international standardization of Enhanced Voice Services (EVS), a high-quality audio-coding technology that is now implemented in all smartphones and adopted by NTT DOCOMO's VoLTE HD+ (Voice over Long-Term Evolution High Definition+) service. I'm currently working on the international standardization of IVAS (Codec for Immersive Voice and Audio Services), which aims to achieve high-quality multipoint telephone services using audio-object coding that enables immersive two-way communication (**Fig. 2**).

In 2020, I served as general chair of Detection and Classification of Acoustic Scenes and Events (DCASE), a major international conference on understanding sound environments, where our team proposed a model for describing the sound on the basis of the distance between language and sound and won first place in the DCASE Challenge Task 6: Automated Audio Captioning task.

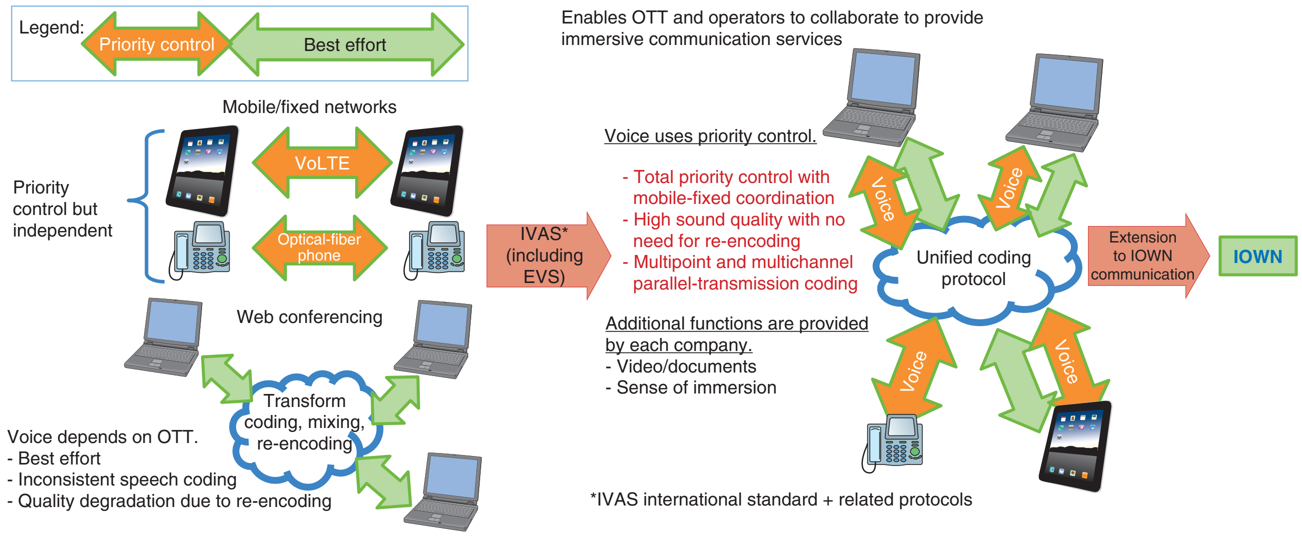
Practical application of basic research results through international standardization

—You have been involved in a wide range of activities from basic research to international standardization. In this context, could you tell us what you value as a researcher?

In my research, I find a research topic (technology) with real-world applications in mind, carefully observe the constraints and requirements, and create a reasonable hypothetical model by comprehensively examining the observation results from physical, psychological, mathematical, and informational perspectives. By using this model, I devise the functions necessary for practical use and the methods to implement them and make proposals to international standardization conferences and organizations toward the goal of our technology being used in society. This model can be created as a precise and complex model or as simple and easy model in accordance with the situation, so I find the appropriate model representation in consideration of accuracy and other requirements for practical use cases.

When considering the requirements from a practical standpoint, for example, in noise detection, I also tried to change my way of thinking. In general, music or voice is a target signal and background noise is an unwanted noise. In anomalous-sound detection, however, it is necessary to construct a model by treating strange noise made by machines as a target signal and normal machine sounds as noise. In terms of how to model the phenomenon of interest, the basic idea is

* MPEG-4 ALS: A lossless compression method standardized as part of MPEG-4 Audio by the Moving Picture Experts Group (MPEG) in International Organization for Standardization/International Electrotechnical Commission (ISO/IEC). It is a data-compression method with which the volume of data before compression and the volume of data that are compressed then decompressed can be completely equal.


Legend:
Priority control
Best effort
Mobile/fixed networks
Priority control but independent
VoLTE
Optical-fiber phone
Web conferencing
Transform coding, mixing, re-encoding
Voice depends on OTT.
- Best effort
- Inconsistent speech coding
- Quality degradation due to re-encoding
IVAS* (including EVS)
Enables OTT and operators to collaborate to provide immersive communication services
Voice uses priority control.
- Total priority control with mobile-fixed coordination
- High sound quality with no need for re-encoding
- Multipoint and multichannel parallel-transmission coding
Additional functions are provided by each company.
- Video/documents
- Sense of immersion
Voice
Voice
Voice
Voice
Unified coding protocol
Extension to IOWN communication
IOWN
*IVAS international standard + related protocols

IOWN: Innovative Optical and Wireless Network
OTT: over the top
VoLTE: voice over Long-Term Evolution

Fig. 2. Aim of 3GPP IVAS standardization at NTT.

the same as the technology used in lossless coding and EVS.

As a corporate research institute, NTT Communication Science Laboratories is engaged in everything from basic research to practical applications, and our desire is to advance research that is useful to society and ultimately contribute to humanity in some form or another. We have therefore set themes that are at the intersection of needs and new technologies or seek to find that intersection. As basic researchers, however, we must also be aware of competition with other researchers. Therefore, there is only so much one person can do, so I believe it is important to focus on what only we can do, rather than trying to do everything by oneself. It is therefore important to work as a team with organic cooperation and division of roles or collaborate with outside parties.

—Is the world of international standardization established through the consensus of participants different from that of basic research?

In the world of international standardization, being useful to society is of paramount importance. Otherwise, there would be no point in establishing an international standard. Therefore, participants in international standardization will make proposals toward this goal and build consensus in the course of discussions. In discussions, conflicts of opinion sometimes arise; even so, the roles of the chairperson, rapporteur, and others, who are basically elected, are to bring those in conflict together and increase the number of supporters to reach a consensus. To fulfill these roles, they need to be trusted by the participants and have broad and deep technical expertise in the field.

My subject technology is only a small part of the international standardization arena, but I still find it frustrating when our technology is not adopted, so I will make every effort to get it adopted. Even if our technology is not adopted directly, I believe that demonstrating it as a competing technology leads to the adoption of easier-to-use and more valuable technology, thereby contributing to humankind.

As for research, each researcher has their own area of expertise and phase of expertise within phases from theme setting to practical application. Even if the themes and goals of different researchers are close, different approaches will result in different time frames. I believe this time-frame difference is one element of the competitiveness in research.

I have a position as both a researcher and contributor to international standardization; however, I am engaged in research with the aim of making practical use of technology, so I am not conscious of each role being a separate entity. As I mentioned earlier, I am always aware of the intersection of needs and new technologies, so issues and requirements discussed in international standardization become needs, which in turn become research themes in terms of pursuing new and basic technologies for satisfying those needs. On the contrary, we may propose the new technology resulting from our research to international-standardization bodies, where discussions can lead to the formation of new methods for applying that technology.

To cite a recent activity, I've been fostering a community at DCASE through the anomalous-sound-detection challenge that I am organizing with other researchers. Through this challenge, we have increased the number of researchers in the field of anomalous-sound detection and have created a platform that enables us to refer to new technologies that emerge from this challenge and eventually gain insights that we can use in our business. Participants in the community are also using this platform to advance their research. Although this challenge differs from international standardization, I believe it will greatly accelerate the process of making a technology useful to society.

Effectively communicating the value of research

—Could you give a message for younger researchers?

In many cases, researchers research a topic that is of interest to them, which can be fun. Such research can be undertaken by convincing those around you—sponsors, customers, partners, operating companies, supervisors, colleagues, and rival researchers—of its value. To get people to understand that value, you need to show them evidence. It is necessary to explain what you aim for, what you want to do, why it is important, and how you will solve the problem then provide evidence to support these elements. However, the people to whom this evidence should be presented—sponsors, customers, partners, operating companies, supervisors, colleagues, and rival researchers—are not all experts, and some are not researchers or engineers. You should effectively communicate the value of your research so that people from various walks of life will understand it. To this end, I believe it is necessary to be aware of the environment surrounding your research.

In the field of basic research, however, disruptive technology that is too advanced to keep up with can suddenly emerge. In such cases, you should not be distracted by what others say; instead, strive forward on the path you believe in. If, as a result of your research, you can demonstrate strong evidence that no one can refute, those around you will eventually be forced to understand. Of course, I think everyone should make an effort to be understood whatever amazing technology you have developed.

■ Interviewee profile

Noboru Harada received a B.S. and M.S. from the Department of Computer Science and Systems Engineering from Kyushu Institute of Technology, Fukuoka, in 1995 and 1997 and later received a Ph.D. from University of Tsukuba, Ibaraki. Since joining NTT in 1997, he has been researching lossless audio coding, high-efficiency coding of speech and audio, and their applications. He is an editor of ISO/ IEC 23000-6:2009 Professional Archival Application Format, ISO/ IEC 14496-5:2001/ Amd.10:2007 reference software MPEG-4 ALS and ITU-T G.711.0 and has contributed to 3GPP EVS.