Researchers Are the Source of Social Progress. Be Confident in Your Research Theme and Aim for the Next Big Topic

Takehiro Moriya
NTT Fellow, NTT Communication Science Laboratories

Abstract

In the wake of the COVID-19 pandemic, remote work, for which image and voice communication using personal computers and smartphones plays a vital role, has become commonplace. NTT Fellow Takehiro Moriya has been engaged in research on speech- and audio-signal coding for about 40 years to improve and innovate communication and quality of life. We interviewed him about the changes in technology development in the field of speech- and audio-signal coding and his attitude as a researcher.

Keywords: speech and audio coding, IVAS, standardization

Pursuing speech- and audio-signal coding

—Could you tell us about the research you have conducted over the past 40 plus years and how it has changed over time?

I have been investigating speech- and audio-signal coding, namely, digitizing voice and music signals and compressing them efficiently and with high reproducibility. For example, the music we enjoy through portable music players and digital broadcasting is not simply digital data of the original signal; it is compressed data in which the volume of the digital data is reduced by about 90%.

Nippon Telegraph and Telephone Public Corporation focused on researching analog speech (including speech quality) transmitted over the limited bandwidth of the telephone and significantly contributed to developing international standards set by the CCITT (Comité Consultatif International Télégraphique et Téléphonique), the predecessor of the International Telecommunication Union - Telecommunication Standardization Sector (ITU-T).

Along with the progress in research on the digitization of telephone networks and terminals, research on the digitization and compression of analog speech signals has also progressed, and this research is the origin of our research on speech- and audio-signal coding at NTT. With the digitization of relay networks, NTT's "INS Net 64" and "INS Net 1500" ISDN (Integrated Services Digital Network) services began in 1984, and as those services became widespread, speech- and audio-signal coding became more important. In the 1990s, second-generation

(2G) mobile phones, which use digital technologies, appeared, and expectation for speech- and audio-signal coding increased thanks to its ability to ensure quality under the severe limitations on transmission bit rates, etc. imposed on mobile networks compared with fixed networks. Our coding technology satisfying certain conditions, such as ensuring sound quality at low bit rates even if a transmission-code error occurs, was adopted in Japan's standard coding scheme for 2G mobile phones following a competitive process. Our elemental technologies were also adopted in 3G mobile phones and Internet protocol (IP) phones, contributing to the improvement in speech quality of mobile telephony throughout the world.

—You have achieved research results that have had an impact on the world.

Speaking of global impact, NTT teams have contributed to several international standardization efforts since the 1990s. The standards to which we have contributed include a speech coding for IP telephony in the ITU-T standards and a low-bit-rate audio coding and lossless coding (which does not allow distortion) in the ISO/IEC (International Organization for Standardization/International Electrotechnical Commission) MPEG (Moving Picture Experts Group) standards. Around 2010, the 3rd Generation Partnership Project (3GPP), an international standardization organization for mobile communication systems, began developing a speech-coding standard. This is because there was a strong need to establish a new speech-coding scheme for Voice over Long-Term Evolution (VoLTE), the speech communication standard for worldwide 4G mobile communication systems. In response, the NTT Group and many other experts from around the world competed and cooperated to develop an integrated speech and audio codec called Enhanced Voice Services (EVS), which became an international standard in 2015.

Until then, speech-coding standard for mobile phones had used the Code Excited Linear Prediction (CELP) algorithm, which emulates the human voice production mechanism, to transmit human voice at a low bit rate and high quality. Combining the CELP algorithm with newly developed low-latency coding modules for music, EVS enabled low-latency transmission of speech (including background noise and music) and music with high sound quality, which had not been possible before. In the process of its standardization, EVS was subjected to large-scale subjective quality evaluation tests under various conditions (with different sound sources and languages) by a third-party organization, and the test results confirmed that EVS achieves much higher quality than that achieved with conventional schemes.

As a result, the EVS codec has been simultaneously adopted by telecommunications carriers, telecommunications-equipment manufacturers, and chip manufacturers worldwide. With the adoption of EVS, the smartphones currently in use around the world can provide high-bandwidth, high-quality calls, regardless of telecommunications carrier or equipment manufacturer. This is the result of a long and repeated process of trial and error by NTT's teams and leading researchers and engineers around the world to improve the sound quality of telephones.

Challenge to establish IVAS standards

—It seems that sound quality is becoming even more important these days.

Due to the COVID-19 pandemic, the number of web conferences and other online activities has increased rapidly. Under such circumstances, NTT Data Institute of Management Consulting and the audio manufacturer Shure conducted demonstration tests to examine the effects of differences in digital audio quality on biological stress reactions during online meetings, and the test results revealed that 85% of users were dissatisfied with the audio quality of web conferencing. The specific complaint is that poor sound quality in meetings causes participants considerable stress as well as prevents them from understanding the content.

Although international standardization set by the 3GPP has improved the sound quality of calls between smartphones, the voice quality in web conferencing is still unstable due to delays and packet loss, which occur because conversations are conducted via personal computers (PCs) and best-effort Internet services are used. The quality of web-conferencing applications for PCs and other devices often deteriorates because processing delays and packet loss are not sufficiently addressed. However, society is increasingly demanding technologies that give the participants in a web conference a "sense of presence" or "immersive experience" as if they were meeting in the same place, including new communication venues such as the metaverse. Although many people think a high-definition image is a major factor

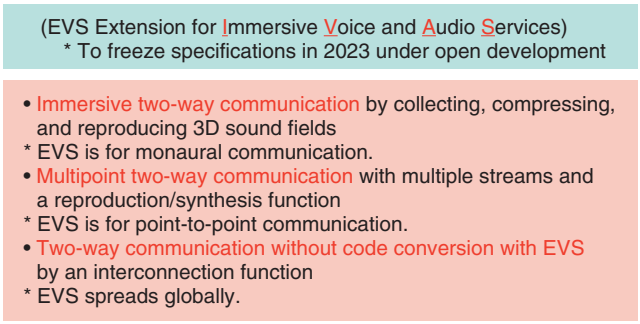

Fig. 1. Goals with IVAS.

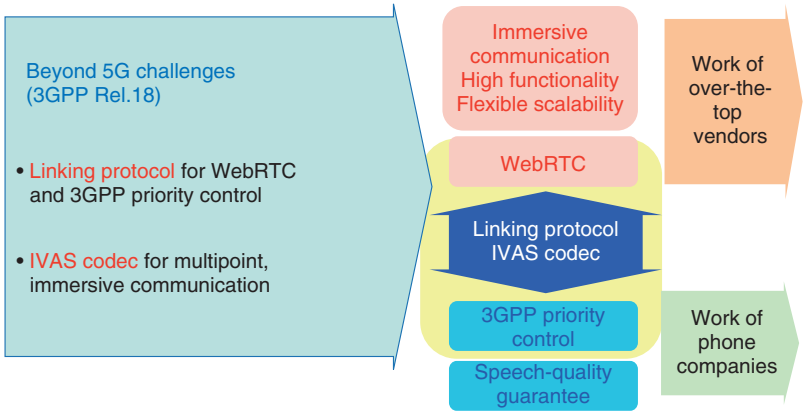

Fig. 2. High-functionality phones with guaranteed speech quality.

in achieving an immersive experience, a great deal depends on the quality of speech and audio. With that background in mind, we are working on standardizing an extension of EVS called Immersive Voice and Audio Services (IVAS). IVAS is under open development to freeze its specifications by the 3GPP in 2023. The aims with IVAS are as follows: (i) immersive two-way communication by collecting, compressing, and reproducing three-dimensional (3D) sound fields, (ii) multipoint two-way communication with multiple streams and reproduction/synthesis function, and (iii) two-way communication without code conversion with EVS by an interconnection function (**Fig. 1**).

The 3GPP aims to (i) ensure reliable communication suitable for voice communication over, for example, optical-fiber (VoIP) and mobile (VoLTE) networks, which have been built by network experts, and (ii) design a protocol that is compatible with Web Real-Time Communication (WebRTC), enabling various functional extensions to be made freely. Therefore, more-stable, higher-quality voice communication than that provided with current web conferencing will be guaranteed, and it is expected to lead to communication with higher functionality and immersive experiences as well as to various new forms of communication, including XR (cross reality) and the metaverse (**Fig. 2**).

—While new technologies are being created one after another, you continue to pursue speech technology, which has a great impact on the immersive experience.

As a company that provides both mobile and fixed-line telephone services, the NTT Group needs to be committed to developing speech-related technologies, and I have been pursuing those technologies with this need in mind. As a researcher at NTT, which has a long history of research into speech in telephony,

I have always wanted to create something that could be used in the business world.

Regarding IOWN (the Innovative Optical and Wireless Network), which the NTT Group is researching and developing to finalize specifications in 2024 and implementation in 2030, speech technology may seem somewhat subdued compared with the glamour of high-speed, large-capacity transmission technology, high-definition-image technology, and so on. However, working in the field of speech, which has a significant effect on our immersive experience, I want to focus on achieving high-quality voice communication regardless of the medium. That is my goal because high speed and large capacity alone do not necessarily improve sound quality.

Research on artificial intelligence (AI) is currently all the rage, and many researchers are in a fierce competition to obtain the best performance. However, only a handful of researchers make it to the top. There are many important fields studied around the world that are not affected by trends. I think one way to be a researcher is to carry out one field of research that interests you. Although research on speech is not a glamorous topic today, it is certainly an important one and is what I chose to pursue. It has many issues that need to be resolved. I want to connect the skills of experts—inherited from my predecessors—to those who will come after me in a manner that paves the way for resolving issues that will not be resolved in my time.

After almost 40 years, AI is currently experiencing its third wave. Thus, a fad focused on one topic does not last very long, but another wave is sure to follow. I think it would be best to find a research theme with an eye on that next wave. In the field of speech, the importance of the digital speech compression and coding technology that I have been researching was reaffirmed in the 1990s with the digitization of mobile phones. Although we should not be complacent, we should be confident in our own research themes and aim for the next big topic.

The pursuit of speech and sound quality contributes to saving lives

—It is important to have confidence in one's own research theme, isn't it?

I believe that themes that can contribute to the world are more important than themes that attract the world's attention or are in vogue. In that sense, for example, the number of fatalities due to traffic accidents has dramatically decreased in Japan since 1995, and that trend must be due to the widespread use of mobile phones, which have made it possible to call an ambulance immediately after an accident. Due to the COVID-19 pandemic, restrictions have been imposed on visiting hospitals and nursing homes. Under such circumstances, many people have been comforted by being able to communicate with loved ones remotely via mobile phones and other means of communication. The spread of mobile phones is partly due to the pursuit of speech and sound quality that we, including our predecessors, have been focusing on, and the results of that pursuit have not only made our society more convenient but also made us realize that the mobile phones held in our hands save lives.

Although some say that the research field of digital compression and coding technology for speech will taper off as communication speeds and capacities increase, it is not unnecessary research. We are currently working on international standardization at the 3GPP as well as research and development with an eye toward Beyond 5G and 6G. Of course, it goes without saying that we aim to contribute to society. As a corporate researcher, it is only natural that I should do what is beneficial to the company. Having said that, I want to work on my research with a smile when I think of contributing to society beyond NTT.

—What do you think a researcher is to society?

I think that researchers are the source of social progress. I often compare politicians with researchers. For example, if 200 people need something that can only be given to 100 people, a politician might consider either giving it to those who really need it out of the 200 or giving everyone a half share of it. On the contrary, researchers take on the challenge of creating value by turning 100 into 200 or even 1000. This way of thinking is exactly what led me to halve the amount of transmitted information by compressing the speech data so as to double the efficiency of utilization of the radio waves when the digitization of mobile phones began.

Although I have passed retirement age, I am continuing my research on speech as an NTT Fellow. I consider my status as proof of the importance that NTT attaches to research on speech. Accordingly, I want to strive to continue research on the theme of speech by explaining to younger researchers the importance and value of the research on speech and by conveying its appeal to them. It is important to maintain the belief that this research is important,

even if society does not yet recognize its importance.

As a researcher, you should be sensitive to research trends, but do not follow fads. While it is important to pursue research in depth, you should not be splitting hairs. Writing papers is important, but it is also important not to end your research with self-satisfaction just because you have written a paper. I hope that young researchers will become long-lived researchers who pursue a theme that they can stay true to while keeping up with trends.

■ **Interviewee profile**

Takehiro Moriya received a B.S., M.S., and Ph.D. in mathematical engineering and instrumentation physics from the University of Tokyo in 1978, 1980, and 1989. Since joining the Nippon Telegraph and Telephone Public Corporation (now NTT) in 1980, he has been engaged in research on medium to low bitrate speech and audio coding. In 1989, he worked at AT&T Bell Laboratories, NJ, USA, as a visiting researcher. Since 1990, he has contributed to the standardization of coding techniques for the Japanese Personal Digital Cellular system, ITU-T G.729, G.711.0, ISO/IEC MPEG, MPEG-4 General Audio Coding, MPEG-4 Audio Lossless Coding, and 3GPP EVS. He is an honorary member and Fellow of IEICE (Institute of Electronics, Information and Communication Engineers), Life Fellow of IEEE (Institute of Electrical and Electronics Engineers), member of the Information Processing Society of Japan, and honorary member of the Acoustical Society of Japan.