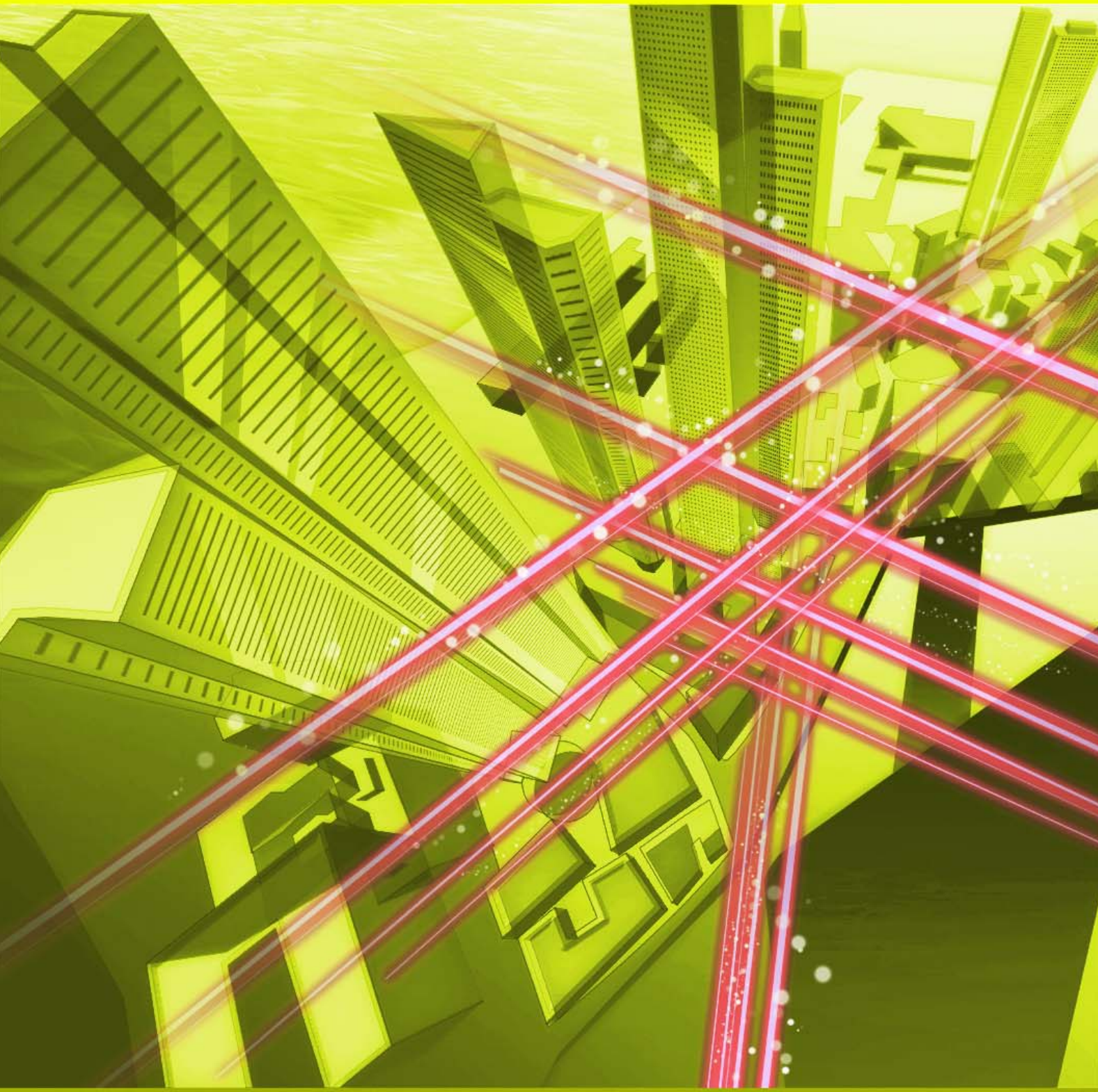
NTT Technical Review

9
2021



September 2021 Vol. 19 No. 9

NTT Technical Review

September 2021 Vol. 19 No. 9

Bridging the Gap among Research, Development, and Business Departments to Achieve a Common Goal

Hidehiro Tsukano
Senior Vice President, Head of NTT IOWN Integrated Innovation Center

Overview

In addition to the three laboratory groups that have been the cornerstones of NTT's research and development (i.e., NTT Service Innovation Laboratory Group, NTT Information Network Laboratory Group, and NTT Science and Core Technology Laboratory Group), the NTT Innovative Optical and Wireless Network (IOWN) Integrated Innovation Center (IIC) was established on July 1, 2021 to extend technology development closer to the commercial implementation stage. IIC is striving to create and implement photonics-electronics convergence technology, which fuses optical and electrical signals and is key to enable IOWN. We interviewed Hidehiro Tsukano, head of IIC, about the purpose of the establishment and mission of IIC as well as the qualities required of top management.

Keywords: IOWN, photonics-electronics convergence, R&D

Realize the Innovative Optical and Wireless Network (IOWN) concept and translate the real world into the digital virtualized world through introducing photonics-electronics convergence technology to contribute building a sustainable society

—You were appointed the head of the newly established NTT IOWN Integrated Innovation Center (IIC). Could you give us some details about the center?

Under the IOWN initiative announced in 2019, NTT aims to implement an innovative network and information-processing infrastructure introducing photonics-electronics convergence technologies by 2030. Tasked with strengthening our research and development (R&D) capabilities toward that aim, IIC began operation in July 2021 by re-assembling the R&D resources of NTT laboratories.

IIC consists of three centers: (i) NTT Network Innovation Center (NIC), which embodies the IOWN concept and is responsible for R&D of innovative network systems that support the integration of mobile and fixed networks; (ii) NTT Software Innovation Center (SIC), which promotes R&D of innovative computing infrastructure for implementing IOWN; and (iii) NTT Device Innovation Center (DIC), which promotes R&D of photonics-electronics converged devices and information-processing

devices for enabling IOWN.

Under this new organization, we will bring together and integrate device technology, network technology, and software-infrastructure technology to develop a game changer in the world of technology and revitalize technological capabilities of Japan. With the spread of information and communication technology (ICT), the global economy has become increasingly borderless, and the meaning of the network and information-processing infrastructure is becoming increasingly important. We believe that it is essential to collaborate with global vendors and will accelerate our R&D by actively engaging with IOWN Global Forum and encouraging development with each vendor.

IIC is an R&D organization, but the name includes "Center," rather than "Laboratory," which I think has a significant meaning. I believe the term "Center" requires a change in mindset to focus on development rather than research. Although NTT laboratories have participated in developing products for practical use, their focus has been more on research. By positioning this development closer to business, we are expected to work with our operating companies to develop business and build products that can be used by players outside the NTT Group. Through these activities, we hope to make a significant contribution to society, which is the NTT Group's primary mission that has been built up over the years since the days of Nippon Telegraph and Telephone Public Corporation.

—I think that to bridge the gap between R&D and commercialization, it is necessary to understand both sides. Have you made use of your experience thus far to do so?

After graduating from university, I joined Fujitsu Limited, where I was first assigned to the purchasing department. Since then, I have been involved in semiconductor and network businesses. As senior executive vice president and chief financial officer (CFO), I gained experience in identifying new business areas from a management perspective with limited resources. I have also built many connections with executive management and key people in global vendors of semiconductors and ICT-related products.

To realize IOWN, it is essential to select various technologies, determine the business feasibility of those technologies, and collaborate with many partners, including through IOWN Global Forum. I believe that my experience can be applied in these areas. The first step in these efforts is to create a *super white box* that enables ultra-low power consumption and tremendously low latency transmission through the application of photonics-electronics converged devices, etc. It will probably take about 10 years to get through, but we hope to accomplish it in that period. We also would like to redirect and reinforce our R&D on Business Support and Operations Support System in an area of Total Operation Management. We want to materialize the vision of IOWN by developing these specific technologies while matching them with the roadmap for IOWN.

Do the right thing right

—You have been at the head of various organizations and offices. What are your thoughts on the qualities required of a top executive?

I think it is important for the people in top management to have beliefs. Of course, such beliefs should not be distorted. By listening to others through various interactions and examining if one's beliefs are truly correct, you can build them into something unshakable. During this process, it is important to be flexible enough to admit when you are wrong and to correct yourself swiftly.

The truth is that there is only one truth, so it is a matter of repeatedly talking to experts and others to understand mechanisms and principles while comparing them to your beliefs and considering plans to make those beliefs a reality.

My belief that I have built up in this way is to "do the right thing right." Although the meaning of "right" may be completely different from different viewpoints, "what we want to achieve" is the same. I therefore investigate how to do something right, listen to people, and deep dive into whether the decision is right or should be implemented. I don't think I became able to do this overnight, that is, it is the accumulation of various experiences that I had over the years.

—It is important to listen to the words of others and be exposed to their knowledge so that you can make the right decision.

I think that top management also needs the ability to communicate clearly and directly. I joined Fujitsu in 1981, and at that time, Japan's national power was on the rise. As far as I remember, after the USA, Japan used to be the world's second largest producer and consumer in information technology and electronics market, but now it has been overtaken by China. I think this outcome shows that Japan has been resting on its past laurels, while others have been growing faster than Japan.

In addition, the so-called "excellent companies" have established their own specialties, have their market share in their countries, and have a large presence in foreign countries. They also have a business-level language skill of the country they want to compete in. In the service field, it is especially important to understand the language and culture of the target country, so my theory is that if you cannot communicate with the people of the country using their language at the closest level possible, you shouldn't compete there. It is not a simple matter of hiring local people; instead, it is important to be able to communicate directly with customers and stakeholders.

I'd say that Japan has cut corners in the area of language skills. I have almost 10 years of work experience in the USA, including half-year business trips. When I was first assigned to the role in the USA as the procurement manager of a plant with about 1000 employees, I was afraid to make phone calls in English, and I wasn't sure if I'd be able to convey what I intended. However, I thought that I couldn't carry out my duties and grow myself if that situation continued, so I made the effort to walk around the vast factory and communicate with the colleagues so that I'd become involved in all the processes under my charge. At first, I found it difficult to communicate with them smoothly, but I kept at it anyway with determination. Once I got a little more familiar with the environment, I stopped worrying about being on the phone, and after about three years, I became more confident. With that confidence, I decided to be positive and try things even if I failed. I adopted that mindset because you can't start anything if you always choose to do things on the safe side. Through such trial and error, I have learned not to be afraid of failure.

Recommendations of John Manjiro

—The establishment of beliefs and smooth communication is rooted in hard work.

My motto, which was instilled in me at my integrated junior and senior high school, is "*shitsujitsu gouken*" (sincere and sturdy). Regardless whether you can succeed in accepting diverse culture, nothing will start unless you take on challenges yourself. I face everything with this attitude. Since I took up a corporate management position around 2010, I have been recommending those around me to learn about John Manjiro*. In other words, going overseas to work. I tell people that if there is something they want to do, they shouldn't just think about doing it only in Japan; instead, they should visit countries and try it there first.

If you have the passion to do something, the other party will meet you at least once and listen to you. To prepare for that opportunity, I think we need to cultivate our resourcefulness on a daily basis. Knowledge in the liberal arts is especially important. Painting, music, and sports are good topics for starting a conversation. The amount of interest you have in the other person may determine whether you can make the effort to cultivate your ability to respond to any topic and gain knowledge. Then, if you can share your beliefs and purpose, the relationship will grow deeper. I have realized from my own experience and the actions of senior management that it is very important to take time to build a mutual understanding and a solid relationship even if the relationship is not directly related to business.

If you have a friend or acquaintance in common with the person you want to do business with, you can ask them to introduce you to that person so you can propose "Let's start something together." These relationships are what we call "communication paths." Nothing happens overnight; therefore, we need to build up these paths from the time we join the company or even earlier. However, we can only know later the results and effects built in this manner. The results through efforts always come later, and to put it bluntly, what we have done can be evaluated and proven only by time and history.

As with all things in nature, time passes equally and fairly to everyone. I believe that it is necessary to constantly think, act, and take an interest in various things to keep improving and evolving ourselves. I think that giving up or neglecting something means that you are not moving forward or you have stopped growing and that you are being left behind the times and degenerating. In other words, it's all about keeping the brain working, and for me, stopping growing and degenerating is frightening.

* John Manjiro: A young Japanese fisherman whose ship was swept out to an island in the Pacific Ocean in 1841, rescued by an American ship, and studied English, surveying, and sailing in USA. He then sailed around the world and returned to Japan in 1851. He became a valuable interpreter/translator when Japan was opening up to the world at the end of its isolation period.

—Do you have any words of advice for engineers and researchers?

I think the purpose of science and technology has been realizing dreams or alleviating fear. Either way, I believe that it is the job of engineers and researchers to fulfill their missions of achieving something that has not been possible before. Another important purpose is to achieve cost reduction so that everyone can benefit from a developed technology. For example, in this day and age, if you want to enjoy the ultimate analog sound of vinyl records, it is said to cost around 30 million yen to prepare the conventional equipment such as vacuum tubes. The quality of the sound is said to be so good that even an amateur can tell the difference in sound quality, but it is not easy to acquire such analog equipment. On the contrary, technological advancement has enabled us to easily enjoy music in a digital format. Engineers are those responsible for evolving technology, enabling something that was not possible, and creating a society in which everyone can benefit from that technology by achieving cost reduction.

I want you to keep in mind that the R&D you are conducting as engineers and researchers can contribute to society and keep working hard every day to make it happen. I also hope that society will understand the efforts of our engineers and researchers. You should be proud of yourself for conducting research to ensure that all people can enjoy the benefits of technology in the real world. In that sense, IIC's major mission is to work toward achieving carbon neutrality to curb global warming.

Interviewee profile

■ Career highlights

Hidehiro Tsukano joined Fujitsu Limited in 1981, where he worked in the purchasing department, focusing on semiconductors. In his career at Fujitsu, he became head of Corporate Planning and Business Strategy Office in 2009 and representative director, senior executive vice president, and CFO in 2017. He also has served as chief strategy officer, in charge of all departments as assistant to the president. In 2019, he became vice chairman of the company. After working as an advisor to NTT Advanced Technology Corporation in 2020 and senior advisor in the Research and Development Planning Department of NTT, he assumed his current position in July 2021.

An Essential Quality of a Researcher Is Persistence. Believe What You Are Doing Will Go Well

Shinji Matsuo
Senior Distinguished Researcher, NTT Device Technology Laboratories and NTT Basic Research Laboratories

Overview

The annual power consumption of datacenters in Japan accounted for 1% of Japan's total power consumption in 2015. As the speed and capacity of data processing and transmission increases, power consumption is steadily increasing, and reducing that consumption is becoming a serious issue. To address this issue, Dr. Shinji Matsuo, a senior distinguished researcher at NTT Device Technology Laboratories and NTT Basic Research Laboratories, is researching and developing innovative technologies for high-density integration of compound semiconductors on silicon substrates to enable photonics-electronics converged integrated circuits. We interviewed him about the progress of his research and his attitude as a researcher.

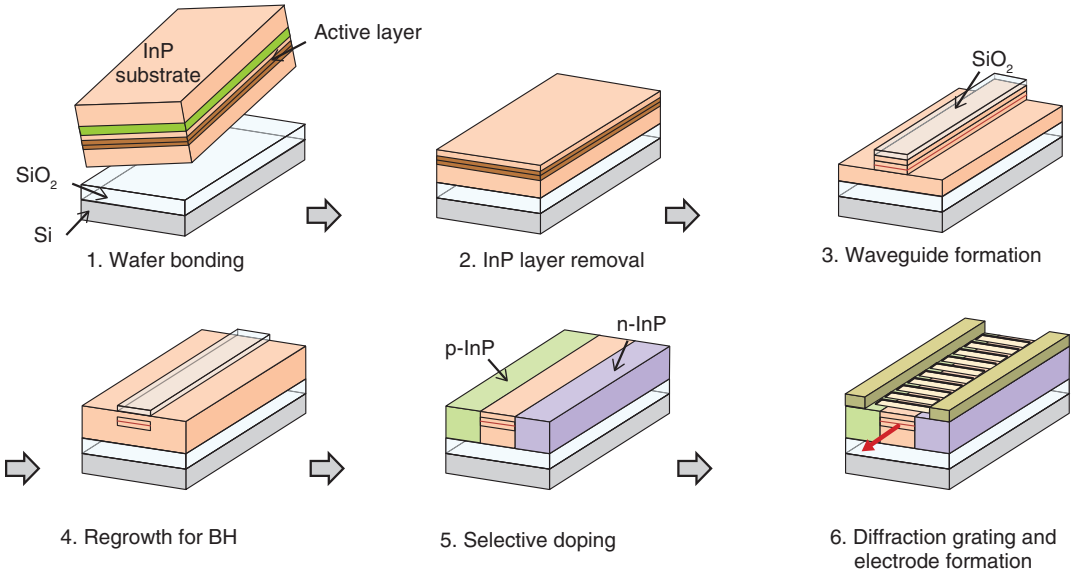
Keywords: photonics-electronics convergence, optical interconnection, directly modulated laser

One aim of the IOWN initiative is to achieve ultra-low power consumption

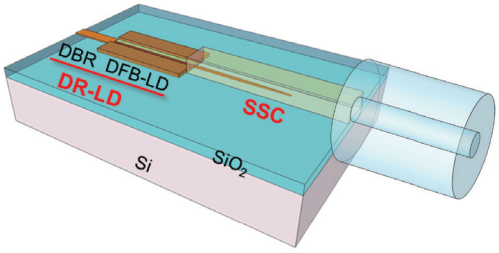
—Would you tell us about your current research?

In 2015, when I was last interviewed, we had just established the basic technology for fabricating semiconductor lasers on silicon (Si) substrates. Since then, our team's efforts have paid off, and we have moved from the research phase to the development phase. At that time, some people were skeptical about our technology because it was new, but now I feel that things are changing favorably. Research on this technology embodies the very characteristics of NTT Device Technology Laboratories, which is engaged in an integrated approach from basic research to applied research on semiconductor devices then to practical application and development of such devices.

We are currently investigating technology for making our devices more socially acceptable and usable, and working on two major themes. One is developing device technology for high-density, low-power, short-range optical interconnections [1]. The development of the Internet-of-Things and the expanded use of artificial intelligence will increase the speed and capacity of data processing and transmission, and power consumption is expected to increase. Therefore, reducing power consumption is a serious issue, and achieving ultra-low power consumption is one of



(a) Fabrication procedure for an LD on Si



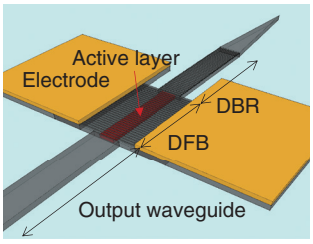
(b) Structure of an LD on Si

BH: buried heterostructures
DBR: distributed Bragg reflector
DFB: distributed feedback
DR: distributed reflection
InP: indium phosphide
SSC: spot size converter

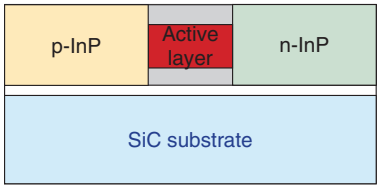
Fig. 1. Fabrication procedure and structure of an LD on Si.

key themes of the Innovative Optical and Wireless Network (IOWN) initiative that NTT has been promoting. To address this issue, we are conducting research and development on optical interconnection to increase the speed and reduce power consumption of electronic equipment used for data processing and transmission by applying optical technology to short-range communication on printed circuit boards of equipment installed in datacenters, etc., which have traditionally been connected by electrical wiring.

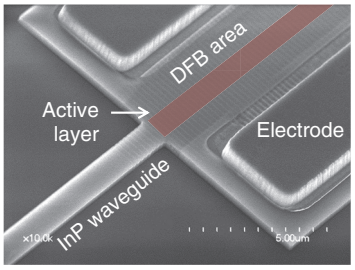
We are developing thin-film (membrane) directly modulated laser diodes (LDs) fabricated on Si substrates as a light source for intra-board optical interconnection (**Fig. 1**). By fabricating LDs on Si substrates, Si photonics technology can be applied to fabricate optical devices, such as wavelength-multiplexing circuits and photodetectors, with high density and at low cost. Forming LDs on a silicon-dioxide (SiO_2) layer having a low refractive index makes it possible to achieve smaller and lower-power-consumption LDs. To fabricate LDs with even lower energy consumption, we are currently developing LDs using photonic crystals. We want to contribute to the development of future information-processing infrastructure by increasing transmission capacity and enhancing high-density integration of LDs as well as integrating LDs with arithmetic processing circuits such as central processing units and graphics



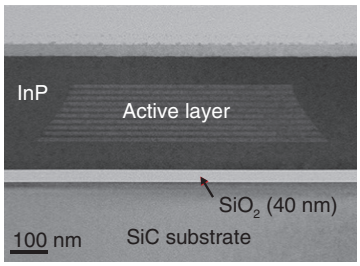
(a) Bird's-eye view



(b) Cross-sectional view near active layer



(c) Scanning-electron-microscope image



(d) Transmission-electron-microscope image of cross-sectional view near active layer

Fig. 2. Directly modulated laser on SiC substrate with bandwidth of over 100 GHz (world's fastest).

processing units.

The other theme is developing a directly modulated laser with the world's fastest bandwidth of over 100 GHz (**Fig. 2**). In collaboration with Tokyo Institute of Technology, we fabricated a membrane laser that uses an indium-phosphide (InP) compound semiconductor on a silicon carbide (SiC) substrate, which has a high thermal conductivity. This laser is the world's first directly modulated laser with a 3-dB bandwidth* exceeding 100 GHz and can transmit signals of 256 Gbit (256 billion bits) per second over 2 km. This achievement was published in the online breaking-news version of the British scientific journal Nature Photonics in 2020 [2, 3].

Although directly modulated lasers are now widely used in datacenters, their limited modulation speed has been an issue. We believe that our technology for fabricating such lasers makes it possible to cope with the expected increase in traffic at low cost and low power consumption, and if research and development of this technology further advances, it will contribute to enabling the high-capacity optical-transmission infrastructure for supporting IOWN. We aim to develop a high-speed, high-capacity communication infrastructure beyond the limits of conventional infrastructure by using innovative technologies centered on optical technology.

—What do you consider important when searching for research themes?

I think it is essential that researchers collect information. I actively participate in academic conferences and talk directly with researchers both inside and outside the company to carefully monitor the trends of NTT's technologies and the technologies that competitors are focusing on. When I was inexperienced, I was sometimes reluctant to attend conferences because I was worried that I would not understand what was being presented or be able to talk to senior or highly respected researchers. However, as I gained experience, those worries have eased. I am now closer to the generation of eminent researchers and speakers and have been invited to give lectures. When I think about the fact that I am now being approached by people to whom I once thought it would be difficult to even speak, I realized the importance of accumulating experience and only experience and effort

* 3-dB bandwidth: A frequency band in which the output power of the laser, which decreases with increasing frequency, attenuates to 70.7%.

make it possible to earn an appreciation from others.

When I was younger, I felt more comfortable studying through papers than interacting with researchers at conferences because I was not used to speaking English. However, I realized that it is difficult to receive new stimuli just by studying in the laboratory and recognized the usefulness of attending a venue such as an academic conference and intensively collecting information and making decisions.

It is also important for your research to be understood by others so that you can continue it. As my interactions with researchers around the world have increased, I have seen them express their ideas in a straightforward and appealing manner, and I have come to understand the importance of expressing ideas assertively. In consideration that research is about creating a brighter future, it is only natural that if you cannot convey to the people in front of you how your research envisions a brighter future, they will not understand it. Moreover, people have an image of researchers being steady and hardworking, and I think we take pride in that reputation; however, considering that we are corporate researchers, I think it is also important to pursue research by thinking about what we can do for our company and society. Therefore, it is necessary to take into account trends such as what our company is focusing on and what issues it is facing.

NTT's laboratories give us annual opportunities to review research plans, discuss them with our group, and promote our plans and learn about the requirements that are being placed on us. While it is important to promote your research to a large external audience, I also want to take advantage of such opportunities to get people close to me to understand the value of my research.

The role of a researcher varies with depth of experience

—Researchers are required to have various skills and roles other than those needed for research, right?

As well as collecting information and building human networks, identifying research themes is an important job for researchers. When identifying a research theme, I consider the personnel, equipment, and facilities and ask myself whether we can bear the responsibilities of using such assets. When it comes to purchasing equipment and other necessities, the financial cost is huge, so a research theme is carefully selected on the basis of whether the pursuit of the theme will benefit society.

On top of the financial costs, research results always come with responsibility. Regarding research on higher-performance semiconductor devices that we pursue, it is important to develop devices and demonstrate the correctness of our fabrication technology. In the development stage, how easily the device can be fabricated is important. However, we should not be bound by these guiding principles. While it is a prerequisite to be able to create something on our own, we also value our sense as researchers that something looks interesting or useful. It is also better to keep in mind that the prospect of "I can fabricate it" differs from person to person.

I believe that the concept of a researcher changes with one's depth of experience. For example, for the first 10 years or so, I thought that a researcher would be able to think of an idea for himself/herself and make that idea a reality on his/her own; however, I later learned the magnitude of the influence of those around me. I am currently a board member and committee member of academic societies and international conferences as well as a senior distinguished researcher at NTT laboratories. Through these experiences, I have come to believe that it is the role of a researcher to think about the research field and the community of researchers in ways like how to create a fruitful research environment for junior researchers and other researchers in the field.

—I heard that you are sometimes asked to give invited lectures at international conferences. Would you tell us about some episodes related to such lectures?

It is a great honor for a researcher to be asked to give an invited lecture. I feel that the invited lectures not only motivate me but also help me self-reflect, reaffirm my position in my research field, and look to the future. My first invited lecture was around 1993, which was five years after I joined NTT. I talked about how to fabricate an LD on a complementary metal-oxide-semiconductor, which is relevant to my current research. At that time, our fabrication technology was highly regarded as innovative, and I was invited to three or four conferences. After I was invited to these conferences, I was desperate not to do anything embarrassing. I had little experience in making presentations in English, and I was given twice as much time as usual to speak, so I practiced hard enough to memorize the presentation.

I have good memories of giving a tutorial-style lecture as a leading expert in the research field at a

conference held in Europe in 2015. I don't know if it was just a coincidence that the atmosphere was like that or if it was normal, but as soon as I took the stage, there was a round of applause. It was an indescribable feeling of elation. In tutorial lectures, experienced lecturers talk about the history of the field and the positioning and vision of their research at the beginning of their lectures. Therefore, I can organize my thoughts by telling my own version of the past, present, and future, and the fact that I can see where I am in the history of my research field and what I am responsible for motivates me to move forward. It is also quite instructive to read through past papers to improve the content of my lecture.

I'm also happy to give lectures to young people. In 2019, I was selected as one of the five lecturers for the Institute of Electrical and Electronics Engineers (IEEE) Photonics Society's Distinguished Lecturer Programs and traveled around and gave lectures in the United States and Canada. In addition to giving lectures on research to graduate students in departments of electronic engineering, I look forward to having the opportunity to interact with students. I'm glad that they listen when I tell them my thoughts on what they should do now. We were also planning to visit India and Brazil; unfortunately, due to the coronavirus pandemic, those trips have been put on hold.

An essential quality of a researcher is persistence

—What would you like to say to young researchers?

Researchers today find it difficult to find time for their research activities, so I fully understand their feeling of impatience. However, in my case, I was able to achieve my current results with a theme launched after I became a midcareer researcher; therefore, I'd like to tell them that they don't have to worry if they don't seem to have enough time or accumulated knowledge.

I work with the belief that what I am doing will go well, and maybe this mindset is what led to my achievement thus far. I think whether you can believe that things will go well affects your motivation. Of course, plenty of studies won't go well even if you believe in them, and I think a large part of success depends on chance and luck. No matter how brilliant you are, some things can go wrong, and some things can go right. I don't know what's causing what, and there is no point thinking negatively, so I'm researching and trusting that it will go well. However, even when it goes well by chance, the sense of not letting go of the opportunity in front of you is crucial. Therefore, it is necessary to collect information. In addition to participating in conferences, actively becoming a committee member or organizer of a conference will greatly expand your opportunities to collect information.

An essential quality of a researcher is persistence. Naturally, research doesn't always go well, and sometimes you will feel pressure from the people around you; however, you can handle that pressure by persistently maintaining your core skills and ideas while being flexible enough to adapt to the times and environment.

Moreover, you should remember that we are members of society. Explain to your immediate seniors and the people around you about what you want to do and convince them of the significance of your research. If you don't have their understanding, you can't continue your research. Then, listen to and incorporate the opinions of those around you as you proceed with your research. It is important to have a sense of balance between assertiveness and acceptance to build a cooperative relationship with the people around us.

One last thing, from the lessons I have learned through my international activities, it is better to polish your English skills from a younger age. If you develop an open mind that lets you talk easily with researchers from all over the world, your future research activities will be more enjoyable. I also believe that casual conversations between researchers in preparation for and after a conference presentation can lead to opportunities for joint research and expand your research activities. Therefore, cherish the friends you make at conferences.

References

[1] K. Takeda, T. Fujii, T. Kishi, K. Shikama, H. Wakita, H. Nishi, T. Sato, T. Tsuchizawa, T. Segawa, N. Sato, and S. Matsuo, "Device Technology for Short-range Optical Interconnections with High Density and Low Power Consumption," NTT Technical Review, Vol. 18, No. 10, pp. 21–30, Oct. 2020.
https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202010fa3.html

[2] S. Yamaoka, N.-P. Diamantopoulos, H. Nishi, R. Nakao, T. Fujii, K. Takeda, T. Hiraki, T. Tsurugaya, S. Kanazawa, H. Tanobe, T. Kakitsuka, T. Tsuchizawa, F. Koyama, and S. Matsuo, "Directly Modulated Membrane Lasers with 108 GHz Bandwidth on a High-thermal-conductivity Silicon Carbide Substrate," Nat. Photon., Vol. 15, pp. 28–35, Oct. 2020.

[3] Press release issued by NTT, "World's Fastest Directly Modulated Laser Exceeding 100-GHz Bandwidth—Membrane Laser on Silicon Carbide Substrate Achieves Low Power Consumption," Oct. 20, 2020.
https://www.ntt.co.jp/news2020/2010e/201020a.html

■ **Interviewee profile**
Shinji Matsuo
Senior Distinguished Researcher, Materials and Devices Laboratory, NTT Device Technology Laboratories and NTT Basic Research Laboratories.

He received a B.E. and M.E. in electrical engineering from Hiroshima University in 1986 and 1988, and Ph.D. in electronics and applied physics from the Tokyo Institute of Technology in 2008. In 1988, he joined NTT Optoelectronics Laboratories, where he researched photonic functional devices using multiple quantum-well p-i-n modulators and vertical cavity surface emitting lasers. In 1997, he researched optical networks using wavelength-division multiplexing technologies at NTT Network Innovation Laboratories. Since 2000, he has been researching high-speed tunable optical filters and lasers at NTT Photonics Laboratories and NTT Device Technology Laboratories. He is an IEEE Fellow and a member of Japan Society of Applied Physics and the Institute of Electronics, Information and Communication Engineers.

Thinking about "Who Benefits?" and "What Is the Benefit?" and Trying Out an Idea as Quickly as Possible

Masayuki Terada
Senior Manager, X-Tech Development Department, NTT DOCOMO

Overview

Technological innovation has made it possible to generate, collect, and store a vast and diverse range of data. There are efforts being made to analyze data that have been overlooked in the past and use them for business. NTT DOCOMO is also using big data from a new perspective to solve social issues. We interviewed Masayuki Terada, who is engaged in the research and practical application of population statistics using mobile network data called Mobile Spatial Statistics and traffic-jam prediction using artificial intelligence called Traffic Congestion Forecasting AI, about the current progress of research and development and the thrill of being a researcher and developer.

Keywords: population statistics, traffic-jam prediction, differential privacy

Research and practical application of technologies for creating, using, and protecting statistical data

—Please tell us about your current activities concerning research and development.

I am currently involved in the research and practical application of the following three themes: creation of statistics from large-scale data, use of statistics for social prediction, and protection of privacy concerning statistics.

Our research and development (R&D) results regarding the creation of statistics from large-scale data are applied to various services. An example is "Mobile Spatial Statistics" [1], which has been used for reporting the increases or decreases in the number of people at terminal stations during the novel coronavirus pandemic (**Fig. 1**). From operational data collected from our mobile phone network, the population distribution of where and how many people are located, by age, gender, and place of residence is estimated hourly and almost in real time throughout Japan. During this estimation process, the privacy of our customers is ensured by removing and aggregating personal identifiers as well as through privacy-preserving processing.

Since 2013, Mobile Spatial Statistics has been used by the government for urban planning and disaster countermeasures and by the private sector for store development and analysis of commercial spheres. When this service was first launched, it did not provide

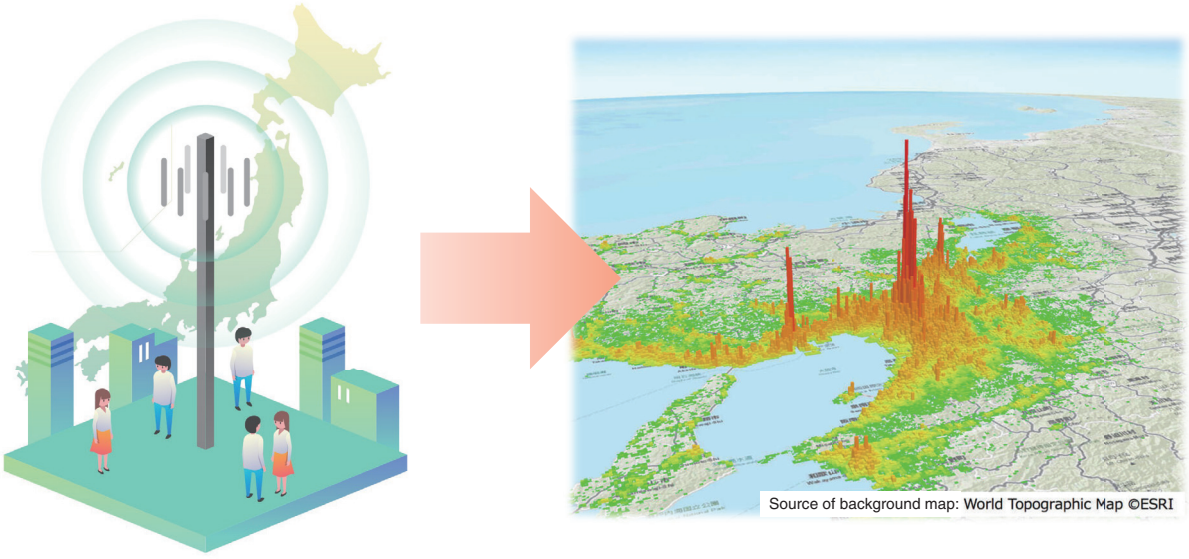


Fig. 1. Mobile Spatial Statistics.

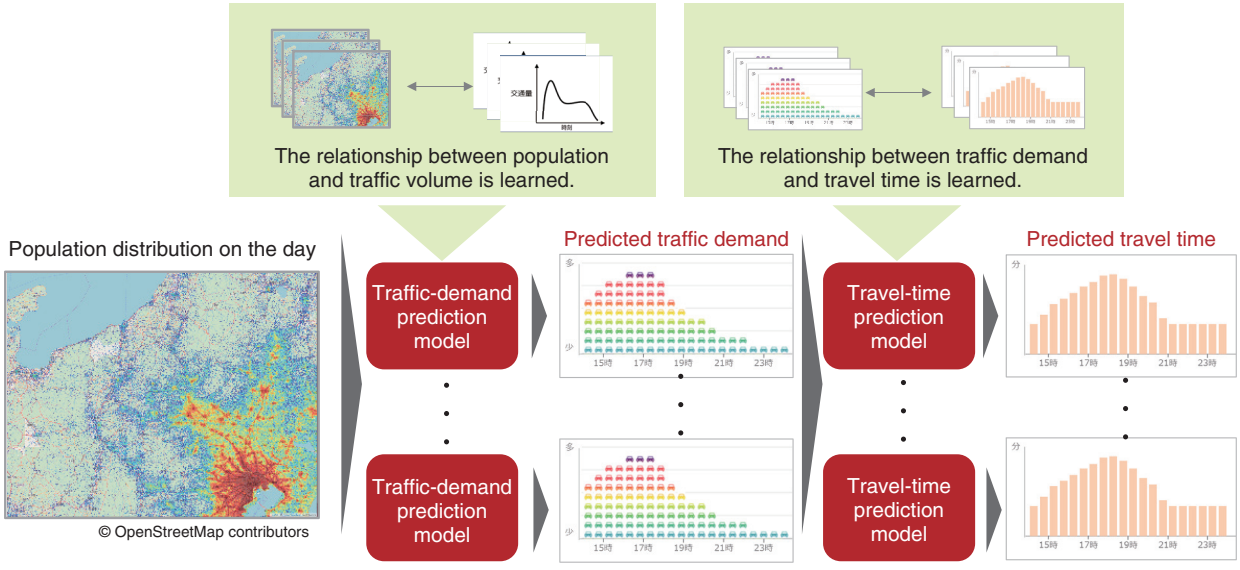


Fig. 2. Traffic Congestion Forecasting AI.

statistics in real time; however, from 2020, it has been providing statistics in near real time in the form of population data from the previous hour or so.

An example of statistics utilization for social prediction is the Traffic Congestion Forecasting AI [2] using Mobile Spatial Statistics (**Fig. 2**). Since social and economic activities are carried out by people, if we can observe the movements of people "now," we will be able to predict the "future" of social phenomena and economic trends. Traffic Congestion Forecasting AI is a world-first application of this principle to predicting traffic congestion. This technology differs from long-term congestion forecasts based on seasonal, day-of-week, and holiday patterns (such as the Japanese Obon festival and New Year's holiday) and short-term forecasts using probe information

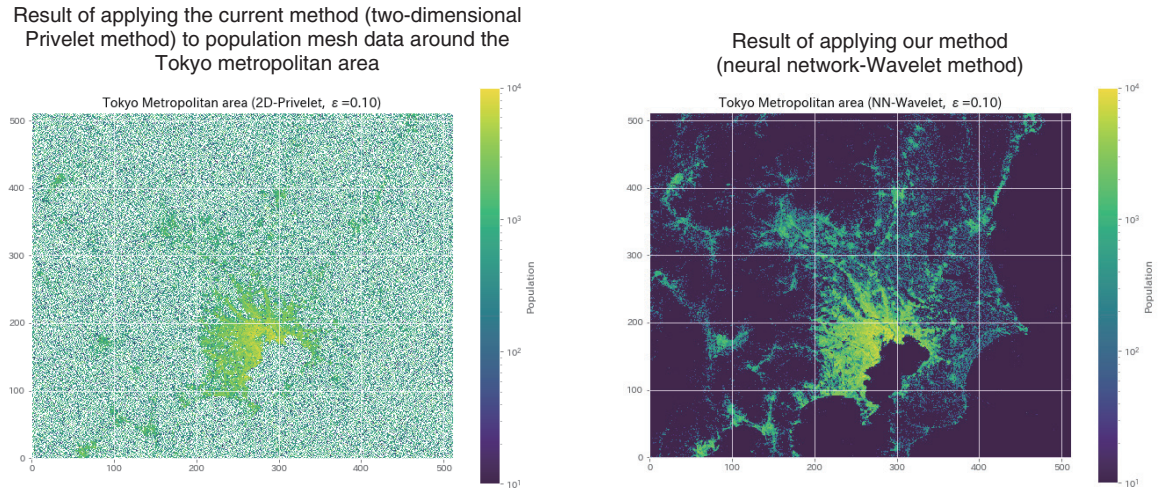


Fig. 3. Applying differential privacy to large-scale tabular data.

from in-vehicle GPS (Global Positioning System), etc. That is, a model is first created by enabling artificial intelligence (AI) to learn sets of past population distributions and traffic volume. Then, the population distribution at a certain time of the day in question is input into the created model, and traffic congestion up to about 10 hours ahead of that time can be predicted.

Traffic Congestion Forecasting AI is being used in a joint experiment with East Nippon Expressway Co., Ltd. (E-NEXCO) to provide traffic-prediction information for the Tokyo Wan Aqua-Line Expressway (a toll way that traverses Tokyo Bay) upstream route (in the Kawasaki direction) and the Kanetsu Expressway upstream route (Numata to Nerima), and that information is distributed daily for widespread use at around 2:00 pm on E-NEXCO's road-information website "Drive Plaza." The use of Traffic Congestion Forecasting AI will make comfortable driving possible without encountering traffic jams. Moreover, if more people use it to avoid traffic jams, traffic jams will decrease or be eliminated through traffic dispersion.

Regarding protection of privacy concerning statistics, we are focusing on applying *differential privacy* [3] to large-scale tabular data (**Fig. 3**). Since the data handled by NTT DOCOMO are often related to personal information, its handling requires strict legal restrictions and social responsibility. Accordingly, privacy-preserving technology, that is, technology that enables data to be used securely while protecting privacy, is essential. Differential privacy is a framework for comprehensively guaranteeing the safety of privacy-protection technologies, and its application is being researched and developed by Google and Apple. In the U.S., it was announced that differential privacy was applied to the 2020 US Census.

To make Mobile Spatial Statistics more secure and useful, we have been investigating how to apply differential privacy to large-scale geospatial data such as Mobile Spatial Statistics and censuses. The results of our research have been reflected in the concept of privacy protection concerning Mobile Spatial Statistics, and we have returned those results to society in the form of academic papers. In 2020, I was appointed as a special professor at the Statistical Research and Training Institute of the Ministry of Internal Affairs and Communications, where I have been investigating how to apply differential privacy for official statistics in Japan.

—How did you find these research themes?

As the saying goes, "Necessity is the mother of invention." If there is a technology that you need but cannot find it in the world no matter how hard you look for it, it will likely become the theme of your R&D. A typical example of mine is applying differential privacy to large-scale data. Since privacy protection is a fundamental issue in regard to Mobile Spatial Statistics, I thought that a strong security guarantee, such as differential privacy, would be needed. However, no matter how many papers I flipped through at the time, I could not find an

appropriate method to satisfy that need. That's when I started to think, "If a method isn't available, I'll develop it myself."

The story is a little different for Traffic Congestion Forecasting AI. One of the reasons for developing it was that I hate traffic jams and want to avoid them. I heard that E-NEXCO is very concerned about the traffic congestion on the Tokyo Wan Aqua-Line Expressway, and I thought that if we could accurately predict the time when traffic congestion would occur, many people, including myself, would avoid that time, and the congestion would be eased. Therefore, I conducted a simple experiment using the real-time version of Mobile Spatial Statistics that I was working on. I was surprised at the accuracy of the prediction results. Therefore, we recollected the data properly and presented the experimental results to E-NEXCO. They showed strong interest, which led to the above-described joint experiment.

However, these stories were only the beginning of our R&D. I owe a lot to the enthusiasm of the people of E-NEXCO for solving traffic jams, the expectations and support of the corporate sales and business departments of our company from the experimental stage, and the constant technical improvements made by my team members and partner companies. I firmly believe that the current form of Traffic Congestion Forecasting AI would not have been possible without their efforts.

Spare no effort to make it easy

—What are the important perspectives in R&D?

The basic research phase, which is conducted to discover new knowledge, and the applied research phase, for creating new products and services, somewhat differ. It is important to determine how applied research results will contribute to society if the research is successful and who is happy with what; in other words, it is important to be able to answer the questions "Who benefits?" and "What is the benefit?"

I think that in many cases, the important mission of R&D efforts in corporations is to create new business and enable profits through the practical application of research results. However, if you jump at a theme just because it looks like it will be good for business, you will tend to follow the lead of other companies. This situation may be an eternal dilemma for those involved in R&D in corporations.

Under these circumstances, it is important to find a winning "one step ahead" theme and focus on it. To find such a theme, the perspective of "Who is happy with what?"—that is, the self-questioning of "Who benefits?" and "What is the benefit?"—is useful. At the stage of searching for a theme, you may not be able to see a concrete market yet, and it may not be clear how to make a business out of a theme; however, if that theme has the potential to make someone happy, in other words, if it benefits someone, you should be able to build some kind of sound business model. Additionally, if you can create a group of people in the business department who are interested in and share the value of a theme by explaining "Who benefits?" and "What is the benefit?" you are aiming for, you can create a profitable business model together with them. I believe that this practice will lead to a broader business than you could have imagined on your own.

—Would you tell us what you have valued as a researcher and developer?

I value the saying "Spare no effort to make things easy." It sounds paradoxical, but I try to think on a daily basis about how I can make things easier. For example, in the case of Traffic Congestion Forecasting AI, if we rely on various types and sources of data for traffic-jam predictions, the design of the system will become increasingly complicated, making it more difficult to improve and maintain; in turn, the implementation and operation of the system will become more difficult. Therefore, we have been designing the system on the basis of using only the population-distribution data of the day and not referring to external data, such as weather information, as much as possible. As we proceeded with our research, we were tempted to use any type of data that could be used to improve accuracy in the immediate future; however, the increase or decrease in the number of people due to external factors, such as the weather, is already included in the population-distribution data, so I asked our team members to be patient and focused on refining the accuracy of the algorithm using the population-distribution data alone.

Partly because of reflections on my past, I try to think carefully about when to put research results into practical use and launch onto the market. I was involved in research on electronic tickets and vouchers in my previous job at NTT's research laboratories around 2000. That research was part of an ambitious project marking the beginning of fintech, and it had both a distribution function of virtual currencies, such as Bitcoin and Etherium, and real payment functions

such as electronic payment. The project was highly regarded internationally and the Internet Engineering Task Force (IETF), which formulates standards for the Internet, established three Request for Comments (RFCs: drafts of standard specifications given by the IETF) in 2003; unfortunately, it was not put to practical use. This project was based on the premise that the world would come to a point where people would always carry an electronic terminal, such as a smartphone, connected to the Internet. However, the first iPhone was not released until 2007, so we had to wait four years after the RFCs were established. The world had not yet caught up with our premise. About 10 years after the establishment of the RFCs, when I had almost forgotten about the project, I was interviewed by a magazine on the theme of "Advanced technology that comes too soon." This experience made me keenly aware of the necessity of identifying current trends, including peripheral technologies, when selecting R&D themes. Since then, we have been trying to present our research results to the world in step with the progress of peripheral technologies and the social environment while simultaneously presenting our future visions.

Keep as many ideas and tools as possible for nimble testing

—You are searching for research themes and taking on the challenge of development while considering various elements, right?

It is not always easy to find a theme that satisfies certain conditions, that is, if a theme matches current trends, one can gain understanding, solve social issues, and respond to our mantra of "Who benefits?" and "What is the benefit?"; therefore, it is necessary to try out a number of different themes. It is said that "Product development is a process of selecting three out of a thousand," meaning only three out of a thousand ideas become reality. And I believe that the key is how quickly you can try out an idea once you have come up with a potentially valuable one. I try to keep as many ideas and tools on hand as possible for this purpose.

I mentioned earlier that our Traffic Congestion Forecasting AI started with a "simple experiment." I was not originally an expert in AI technologies, such as machine learning; even so, machine-learning technologies had been commoditized, and easy libraries had emerged, and my experience of playing around with them to see how easy they were to use came in handy. If I hadn't had that experience or known that such libraries existed, I wouldn't have had the motivation to "give it a try," and maybe Traffic Congestion Forecasting AI wouldn't be around now.

The experts I meet at conferences and lectures, as well as the customers I meet through requests from corporate sales colleague to accompany them, are also very helpful in finding themes and tuning the direction of our R&D. For example, when I talk to people about Mobile Spatial Statistics, it seems that those who are seriously considering using it tend to place more importance on the reliability and accountability of the statistics than on the numerical specifications. To meet these expectations, we are enhancing the brand image of Mobile Spatial Statistics by not only pursuing higher specifications but also on the basis of reliability and security.

—Please give a few words to our junior researchers and developers.

There is a saying, "Standing on the shoulders of giants." It means that our own R&D is based on the accumulation of the results of our predecessors. I think the most exciting part of working in R&D is to be able to contribute to further development of society in the future by putting even a thin layer of your own achievements on top of those of your predecessors. Focused on the three themes that I am currently working on, i.e., creating, using, and protecting statistical data, I hope to put a thin layer of skin on the shoulders of giants and make them as tall as possible. Moreover, I'd be very happy if someone comes along in the future who can stand on top of these giants and make them even taller.

Having said that, I realize that finding a new R&D theme can be difficult. It's not as if you can find a good theme if you moan a lot and think hard. Most of you probably have some kind of theme that you are working on now. Although it is important to thoroughly finish with a theme and release the results to the world, after you finish with the current theme, you may not find something interesting right away. Therefore, in addition to working on your current theme, I encourage you to take the time to talk with people in various fields, and if you find something interesting from those talks, look into it, become involved in it, and try out your ideas. Even if that process doesn't directly result in a new theme, you can gain experience, and if you accumulate a large amount of such experience, it may help you find a theme for the future or give you a hint to solve a difficult problem

concerning your current theme from a different perspective.

Your boss and seniors have probably had the experience of finding new themes in this way, namely, experimenting and playing around with ideas under the umbrella of the current theme. While drawing on their knowledge, I hope that you will also value the pleasurable aspects of your work outside your core business of R&D.

References

[1] Mobile Spatial Statistics (in Japanese), https://mobaku.jp
[2] E-NEXCO Drive Plaza, "Demonstration Tests of Traffic Congestion Forecasting AI" (in Japanese), https://www.driveplaza.com/trip/area/kanto/traffic/ai_traffic_prediction.html
[3] M. Terada, "What Is Differential Privacy?", Systems, Control and Information, Vol. 63, No. 2, pp. 58–63, 2019 (in Japanese). https://www.jstage.jst.go.jp/article/isciesci/63/2/63_58/_pdf/-char/ja

■ **Interviewee profile**
Masayuki Terada

Senior Manager, X-Tech Development Department, NTT DOCOMO, INC.

He received a B.E. and M.E. in engineering from Kobe University, Hyogo, in 1993 and 1995 and Ph.D. in engineering from the University of Electro-Communications, Tokyo, in 2008. He joined NTT in 1995 and moved to NTT DOCOMO in 2003. He is currently engaged in R&D on technologies for privacy protection, population estimation from mobile phone networks, and traffic prediction from population data. He received the IPSJ Outstanding Paper Award and the IPSJ Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) in 2015. He also received the IPSJ Industrial Achievement Award in 2019. He is a member of IPSJ and the Institute of Electronics, Information and Communication Engineers.

Research on Asynchronous Distributed Deep Learning Technology—Optimizing Machine Learning Models in the Age of Distributed Data Storage

Kenta Niwa
Distinguished Researcher, NTT Communication Science Laboratories

Overview

While modern deep learning often requires aggregating data into a single datacenter to train models, in the near future data will be distributed due to increased data volume and privacy protection concern. In this article, we spoke to Kenta Niwa, a distinguished researcher working on asynchronous distributed deep learning technology. This technology allows us to optimize machine learning models as if the data was aggregated in a single datacenter, even in the modern era of distributed data.

Keywords: asynchronous distributed deep learning, machine learning, decentralized system

What is asynchronous distributed deep learning technology?

—Please tell us about your research.

Deep learning is often used for e.g., speech/image recognition. Modern deep learning involves overconcentration of data, namely, all the data is aggregated in a single huge datacenter and then used to train the model. However, considering several industrial applications, such as self-driving vehicles, factory automation, distributed power-grid, and highly personalized models, the volume of data will continue to increase and it will become much more difficult to collect, process, and deploy all the data in a single datacenter. Data aggregation is also becoming more difficult from a privacy perspective due to the effects of legislations, e.g., GDPR (the European Union's General Data Protection Regulation). Because of these factors, we believe that data storage and inference processing will be carried out in a distributed manner in the near future.

We are conducting research to develop an advanced algorithm to virtually manage data sets stored on multiple servers. For example, we recently achieved results with technology that allows us to train

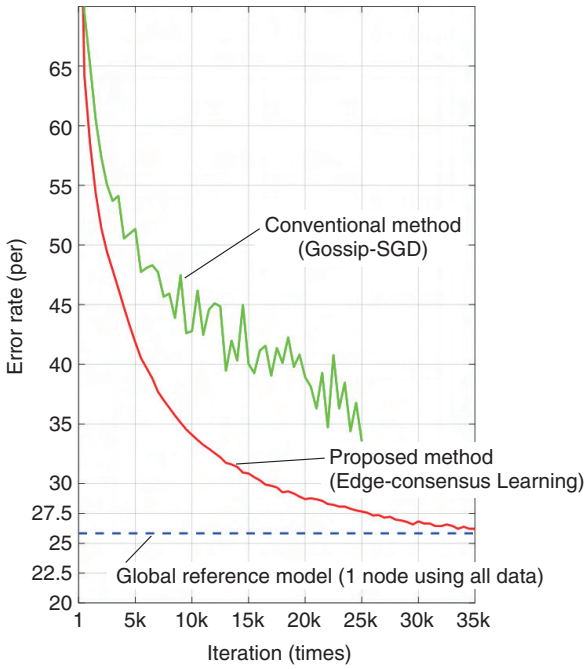


Fig. 1. Comparison among conventional and proposed methods and the global reference model.

machine learning models as if the data was concentrated in one place, even data sets are not aggregated.

—What kinds of methods are specifically being used?

Today, distributed deep learning often uses a method of exchanging and averaging models between servers (forming an average consensus). The formulation of an average consensus is a very simple and effective operation. However, while this method works well if each server has statistically homogeneous data, training often fails to progress if the data on each server is statistically biased (heterogeneous). In addition, as the number of servers increases, synchronous communication becomes more and more difficult.

We built a deep learning algorithm where all the servers connected to the network communicate asynchronously and collaboratively to train the model. I'll spare you the details of the formulas and algorithms, but to put it simply, the research expresses the idea of "multi-node collaboration" in mathematical formula.

For example, it's easy for people who get along well with each other to come together, talk it over, and draw a conclusion. However, if a group of people with strong personalities who don't get along together, everyone will go in different directions. There's not really any point in performing averaging in a situation like this. Think of this algorithm as a way to skillfully express how well the individual elements work together.

Figure 1 shows a comparison between the proposed method (Edge-consensus Learning) and the conventional method (gossip-based stochastic gradient descent (Gossip-SGD)) using a data set commonly used for testing an image classifier called CIFAR-10, with the data distributed in eight servers with statistical bias. The vertical axis represents the classification error rate, with smaller values meaning better performance. The blue dotted line shows the performance of the global reference model that trained using all data collected in one server, the solid green line shows the performance of the conventional method (Gossip-SGD), and the solid red line shows the performance of the new method being proposed (Edge-consensus Learning).

While performance does not reach to the global reference model when using the conventional method, the proposed method gets closer to the performance of the global reference model as learning progresses. I believe we can take this to mean that we have obtained a model more suited to the entirety of the data, even with asynchronous communication.

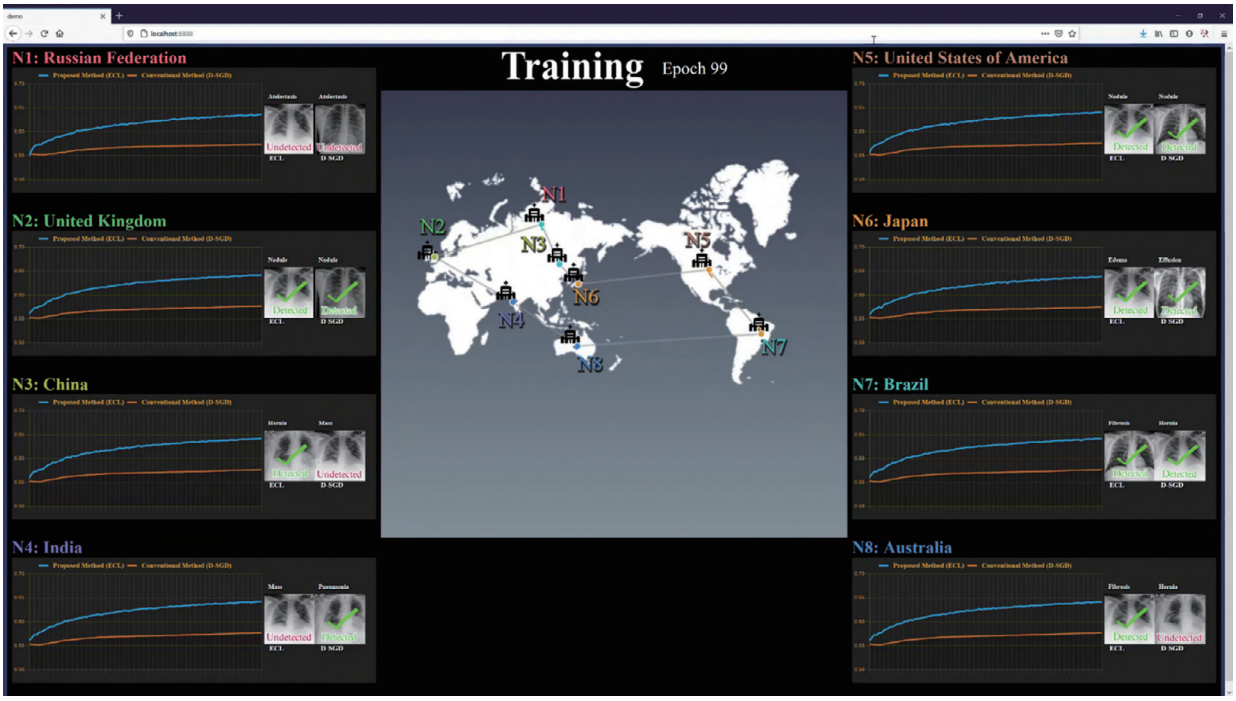


Fig. 2. Combined training of medical image analysis model across multiple hospitals.

—What kinds of possibilities does this research unlock?

Figure 2 shows a demonstration of medical image analysis model using data from multiple hospitals. Medical data is the most important example from the perspective of privacy concerns, and in practice it generally should not leave the hospital. And taking it out of the country is next to impossible. So we considered a network connecting eight hospitals (N1 to N8), where we train the model without taking the imaging data out of the hospital. Specifically, this model is a medical imaging diagnostic aid that uses chest x-ray images to detect the presence of a disease and identify the disease from among 14 different diseases. In light of the differences in diseases handled by hospitals in the real world and regional differences in conditions, we encounter a situation where the number of data items at N1–N8 and the diseases being recorded are statistically biased.

The blue lines show the recognition performance using new model at each of N1 through N8, and the orange lines show the progress of the conventional training method. The medical image data spread across eight hospitals works together well to create a highly advanced level of knowledge, like some kind of "super doctor."

Asynchronous distributed deep learning technology allows us to reap the benefits of machine learning without sacrificing privacy. This technology is not particularly application-dependent, so the algorithm is highly flexible and can be used in various applications such as text translation using smartphone data, training autonomous driving models, voice recognition in call centers, and anomaly detection in industrial factories.

Providing services that leverage NTT's locational advantages

—What do you consider NTT's strengths to be?

NTT has communication stations throughout Japan, and each station is connected by a network. Why not use this locational/physical advantage to install servers in each station to store and process data? Of course, each station will have very different types of data stored, and the system as a whole will be huge. Leveraging this data to bring services such as data processing closer to users is likely to be a viable option for NTT's business in the future. It's what's known as "edge computing." When we started thinking

about this, we thought it might be useful to provide asynchronous distributed machine learning services.

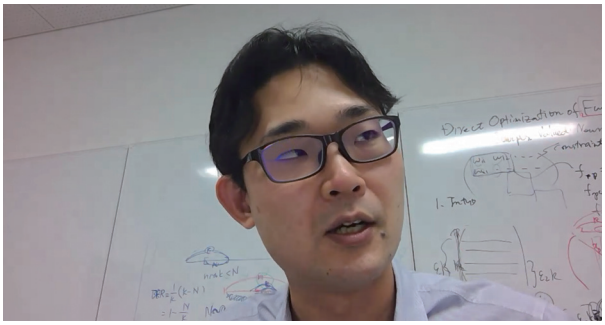
—What do you think about the future direction of your field?

I was originally conducting research related to acoustics, especially communication using microphones and loudspeakers. About 10 years after joining the company, I started studying machine learning after studying abroad in New Zealand, particularly the theme of distributed optimization field.

Many modern systems are constructed in a centralized manner. This is because of the amount of data that can be handled in one place, and because what is needed is considered to be universal, so the same service is distributed to all. However, I believe that services will continue to become more personalized as time goes on. I think we will also offer training and inference for models tailored to individual customers. With this reality, it begs the question "Is the current centralized system still the right one?" It's only natural to consider decentralized manner.

When you think about it, our brains are also "decentralized" in a manner of speaking. We can talk about one thing while thinking about something completely different. I don't think my own mind is doing heavy computing like deep learning and inference in the areas that aren't already burned out! It feels like the distributed computing groups in our heads are connected asynchronously, and they can perform high-level tasks by combining a lot of very light calculations. In this regard, I believe the systems that support the next generation of communications and society as a whole can also be processed at an advanced level with low power consumption and high flexibility, and be durable enough that they won't break even if they get a little damaged.

For example, the IOWN (Innovative Optical and Wireless Network) concept discusses concepts such as traffic coordination of self-driving vehicles in a large-scale smart city in the context of optimizing society as a whole. Of course, every car has a different destination, so I think they should also be driven differently. That said, working only for one's own benefit is no good, and I don't feel like averaging is particularly desirable either. I don't think there are systems that can coordinate and control every car yet, but I would like to create decentralized systems and core software that contribute to these things, keeping "distribution" in mind as a theme.



—What would you say to anyone hoping to become involved in basic research in the future?

The field of machine learning is incredibly competitive and changes every few months. New papers are coming out nearly every day. To be more specific about the intense competition in distributed systems, I think that there is too much competition in distributed learning, but I don't think that's the case when talking about asynchronous distributed systems like the one I'm proposing, which can flexibly obtain a high level of knowledge. Centralized and end-to-end are mainstream right now, and asynchronous distributed systems are still relatively unexplored, and I want to see the possibilities there.

Opening up new fields requires energy. It's easy for an expert on the topic to say "This is how we should do it," but 99.9% of people aren't experts, so it's incredibly important to find the challenges we have in common and discuss, refer to and compare them as we compete with each other. On the other hand, it's also important to create your own unique niches in your policy. I think it's important to work on the things you want to express and build gradually over time.

Nowadays, anyone who can gather data (even high school students) can create models for whatever application they want. In this reality, many of the researchers around me are also struggling to decide whether to focus themselves on basic research or practical development. I have had the opportunity to study abroad, and I learned a great amount about distribution there, so I ended up focusing on basic research. I'm sure I could just as easily have chosen the opposite. I can't say which is better as some people are particularly suited or unsuited to a specific position, but I think in times like these it's best to focus on one end of the spectrum.

■ **Interviewee profile**
Kenta Niwa
Distinguished Researcher, NTT Communication Science Laboratories.

He joined NTT in 2008, engaged in research and development on sound recording processing at NTT Media Intelligence Laboratories. Achieved results in contributing to commercialization of microphone array based speech enhancement technology and the "zoom-in microphone" that can pick up sound clearly from far away. After studying abroad at Victoria University of Wellington in New Zealand from 2017 to 2018, he began research into machine learning, including distributed optimization, at NTT Communication Science Laboratories. He is currently focusing on asynchronous distributed deep learning technology. He also works at NTT Computer and Data Science Laboratories.

Reach Out and Touch Someone's Heart: Exploring the Essence of Communication to Create a Spiritually Rich Society

Takeshi Yamada

Abstract

NTT Communication Science Laboratories has been exploring the essence of communication since its founding 30 years ago. With the aim of achieving communication that *reaches the heart*, its researchers have been creating innovative technologies that approach and exceed human abilities in fields such as media processing and data science. They have also been discovering basic principles that lead to a deeper understanding of humans in fields such as cognitive neuroscience and brain science. This article introduces key activities at NTT Communication Science Laboratories in pursuit of the essence of communication with a look back at past research.

Keywords: artificial intelligence, machine learning, cognitive neuroscience

1. Introduction

This year marks the 30-year anniversary of NTT Communication Science Laboratories (NTT CS Labs), which was founded on July 4, 1991. Throughout these 30 years, it has been exploring the essence of communication and conducting basic research to enable communication that *reaches the heart* [1]. The essence of communication is inherently multifaceted. In addition to (1) conveying information accurately and efficiently, it includes the (2) deepening of mutual understanding by sharing the meaning of information and (3) sharing of underlying intent and emotion by devising creative methods of conveying information, enabling the (4) creation of a spiritually rich society by fostering *heartfelt* contact. With a focus on these four viewpoints of communication, this article introduces important activities at NTT CS Labs in pursuit of the essence of communication while taking a look at past research.

2. Basic technologies for information transfer and speech coding

NTT has been continuously engaged in the research of basic communication technologies such as audio and voice processing and natural language processing since the Nippon Telegraph and Telephone Public Corporation era. The root of this research is speech coding technology, which is one of the most important technologies from the viewpoint of transmitting information accurately and efficiently. Line Spectrum Pair technology proposed by NTT in 1975 is still used in most mobile phones throughout the world as an international standard, and in 2014, it was recognized as an IEEE (Institute of Electrical and Electronics Engineers) Milestone marking a historic achievement in the field of telecommunications [2]. Inheriting this legacy, NTT CS Labs became a leading contributor in the development of Enhanced Voice Services (EVS) technology that was approved as a 3GPP (3rd Generation Partnership Project) standard in 2014. In Japan, this standard has been used since 2016 as

fourth-generation coding in a coding-transmission system shared by three mobile phone companies, and as of 2021, it has been used in smartphones throughout the world. NTT CS Labs has also been contributing to the EVS extension for Immersive Voice and Audio Services called IVAS. More recently, it developed Bitplane Rearrangement for Audio and Voice Encoding (BRAVE), an audio and voice codec featuring robust bit-error performance and low latency. BRAVE was adopted in wireless microphones commercialized by TOA Corporation in February 2021 [3].

3. Obtaining categories and concepts to share meaning

Humans learn by grouping similar things into categories, which enables advanced cognitive activities such as thinking, inferring, decision-making, and communication. It also makes learning itself more efficient. For example, when catching site of an animal, if its form happens to be sufficiently similar to a category that one has already learned, say "cats," that animal would be recognized as a member of that category, in other words, as a cat. While it is difficult to individually remember all objects one has seen up to the present, they can be remembered in a compact form by grouping them into categories. Later, when looking back, it may not be possible to remember the detailed features of that cat, but the fact that it was a cat will not be forgotten. In addition, if something that one encounters is different from any category that one has already learned, one can simply create a new category. Therefore, learning can be made efficient by flexibly increasing, or even decreasing, categories as needed in accordance with data characteristics even for large volumes of data.

NTT CS Labs has been working on achieving such flexible human-type category learning on computer. For example, the products that each customer has purchased can be recorded in matrix form as a history of purchased data that can be used to categorize both customers and products. This type of categorization corresponds to rectangular partitioning of a purchased data matrix. An efficient learning technique based on a Bayesian nonparametric model was proposed that adjusts the optimal partition in accordance with the given input data from among an infinite number of combination patterns in rectangular partitioning [4].

Thus, "cats" as a category is not a specific "cat" but rather an abstraction of "cats" in general. A concept, on the other hand, is a mental representation of a category stored in memory. In other words, it is a set of information that a category points to and consists of what is known about that category [5]. The concept of "cats" that humans hold is not limited to the shape or form of cats. It is rather an integrated abstraction of various aspects of cats, such as the sounds they emit (meowing, etc.), their behavior, the feel of their fur, etc. and the language used to express such aspects. That is, a concept can be acquired by seeing a thing (its set) from different viewpoints (different types of media information or modalities) and be understood as abstract information independent of any viewpoint and expressed as coordinates in a common conceptual space. NTT CS Labs is researching the autonomous acquisition of concepts without having to train a system with correct answers. This can be done by focusing on the co-occurrence of different types of media information such as images and sounds of cats, that is, by using the fact that different types of media information originating from the same thing appear not in a random manner but with specific relationships [6].

4. Communication and language acquisition in infants

Do human infants autonomously learn from the co-occurrence of phenomena in the natural world? To comprehend the essence of communication, NTT CS Labs has been examining communication and language acquisition in infants. For infants, communication is an important means of recognizing objects and promoting the acquisition of knowledge, concepts, and vocabulary. Infants accumulate various types of knowledge from information obtained from the surrounding environment, such as by listening to a parent's conversation, speech from a television, etc., and learn groups of syllables that co-occur with high frequency as words based on statistical learning. However, this does not mean that the infant indiscriminately processes a huge amount of information. Research conducted at NTT CS Labs has found that learning in infants is promoted by communication signals from a parent such as utterances directed toward the infant [7]. The infant uses such communication signals as a learning cue to focus appropriately on learning targets and sort out what to learn from the environment and how.

As explained above, parent-infant communication promotes brain development of the infant. It further affects subsequent vocabulary growth. NTT CS Labs

has been promoting research on language acquisition in infants, and from the results of that research, more than 280,000 copies of picture books for young children supervised by NTT CS Labs have been published. What is significant is that these books are not digital but rather printed material that children can interact with using the five senses. More recently, "personalized educational picture books" was proposed in collaboration with NTT Printing Corporation. These are educational picture books with pictures emphasizing new words for an individual child to learn on the basis of vocabulary checking conducted by the child's parents and on a child-vocabulary-development database developed through research at NTT CS Labs. This venture began with picture books to be read out loud to children, but more recent research at NTT CS Labs revealed that an understanding of characters and their correspondence to sounds actually starts around three years old, slightly before the ability to read and write *hiragana* (Japanese syllabic characters). Thus "names-in-*hiragana/katakana* picture books" was proposed to generate interest in characters targeting children of about three years old. Personalized educational picture books can now be ordered online at ehon.nttprint.com [8].

NTT CS Labs has also undertaken research on the interaction between a parent and infant focusing on the parenting side. Parenting stress and postpartum depression in mothers, child abuse and neglect, etc. have become problems throughout society. To study how mothers interact with infants, types of infant vocalizations and the manner in which a mother approaches her infant in response to those sounds were investigated. It was found that a mother would reflexively respond to the sound of crying and that her response in approaching her infant would become stronger the more urgent those sounds feel to her. In short, the sound of crying arouses a feeling of wanting to respond quickly (a sense of urgency) in the mother, who then approaches the child in a reflexive, unconscious manner.

Humans are equipped with a mechanism for suppressing this reaction. Specifically, there is a hormone called oxytocin. This hormone serves to secrete mother's milk. It is also called a prosocial hormone since it is known to easily arouse positive emotions with respect to another person. The concentration of oxytocin is also known to have a positive correlation with caregiving motivation in the mother. Research at NTT CS Labs has found that oxytocin suppresses this reflexive impulse in the mother to approach her infant at the sound of crying [9]. This research suggests that if the level of oxytocin is low, the mother loses her composure and wants to quickly stop her baby from crying, but if the level of oxytocin is high, parasympathetic nerve activity increases, resulting in making the mother more relaxed and suppressing a reflexive approach to her crying baby. This result may lead to knowledge on how to promote a sense of well-being in parenting.

5. Creating new forms of communication

NTT CS Labs is also working on ways of communicating, that is, on creating new forms of communication. "The medium is the message" is the famous phrase coined by Marshall McLuhan, a scholar of English literature. Through these words, McLuhan asserts that a message includes the means of conveying the message and stresses the importance of the medium that transmits the message in communication, that is, the sensory image that the medium itself possesses. In this regard, there is the famous slogan "Reach out and touch someone" used by AT&T, the American telecommunications company, in commercials in the 1970s with the intention of softening its stiff image [10]. McLuhan contributed to the creation of this catchphrase, which was novel for its time.

In line with this "reach out" point of view, NTT CS Labs once researched a room-sized remote communication system called "t-Room." This is a system in which multiple geographically and temporally separated users share the "feeling of being in the same room" while being at remote locations [11]. However, t-Room did not include the sharing of the sense of touch. For this reason, NTT CS Labs is now researching new sensation-presentation technology using touch that would enable *kansei* (emotionally rich and sensitive) communication to convey deep feelings by touch. The "Mega-Futuristic Experiential Public Telephone" (versions 3 and 4) proposed in 2018 is a touch-based communication system in which pressing the push buttons of a telephone causes a variety of tactile sensations to stimulate the other party's body. More recently, new systems such as Remote High Five and Public Booth for Vibrotactile Communication that truly share tactile sensations beyond distance have been proposed [12].

NTT CS Labs is also researching a means of speech conversion that would enable content that one would like to convey to be freely converted to one's desired form of expression for transmitting and receiving. This research is expected to create new forms of

communication that extend human vocal and auditory functions.

To achieve communication that *reaches the heart*, methods are needed for picking up what a person is feeling from the outside without placing too much of a burden on that person. Regarding the well-known saying, "the eyes are the window to the soul," NTT CS Labs discovered that a human's pupil unconsciously constricts on seeing an attractive face. Consequently, if the size of a person's pupil can be measured in real time, it would be possible to pick up what that person is feeling to some extent. At the same time, it was found that making a person's pupil constrict through luminance/contrast changes actually enhanced the attractiveness of a face as seen by that person. This result suggests that controlling—as opposed to measuring—the size of a pupil could change to some extent that person's preferences [13].

6. Spiritually rich society with diverse values in harmony

Finally, from the viewpoint of *heartfelt* contact, I would like to introduce research in pursuit of the essence of communication from a slightly different angle. It is often said that modern society is becoming increasingly divided. Due to the flood of information, people are becoming increasingly confrontational when they see two sides of an issue that appear to be at odds with each other. Instead of listening to different opinions, they take one side while sacrificing the other, as in globalism or nationalism, centralization or decentralization, and analog or digital. However, precisely for this reason, it is indispensable to create a spiritually rich society that allows for contradictions, recognizes diverse values, deepens mutual understanding through communication while protecting privacy, and nurtures *heartfelt* contact by promoting empathy.

In machine learning, especially deep learning, a massive increase in data and the need to protect privacy are generating a need to distribute and store training data on a group of local servers. However, if each server should train locally under these conditions, the end result will be trained models that are different and mutually contradictory. These models will not converge if they are poorly coordinated. Therefore, NTT CS Labs devised an asynchronous distributed deep-learning framework in which data dispersedly stored on a group of dispersedly located servers can be trained as a global model as if the data were consolidated at one location by having the servers communicate with each other to build a consensus [14].

NTT CS Labs also devised three-dimensional (3D) video-generation technology to enable clear viewing of the corresponding 2D image with the naked eye. In short, this is technology that enables people who prefer to view 3D images with 3D glasses and people who are uncomfortable with 3D and prefer 2D to enjoy the same image together without sacrificing the needs of the other [15].

7. Conclusion

In this article, I introduced key activities at NTT CS Labs in pursuit of the essence of communication with the aim of achieving communication that *reaches the heart*, or more exactly, that *reaches out and touches someone's heart*. The way we interact with other people is changing due to countermeasures and restrictions related to the COVID-19 pandemic. It is especially necessary at this time to pursue the possibilities of new media leveraging the five senses while identifying and solving any problems that may arise in this pursuit. Before trying to convey one's feelings to someone far away, one should realize that people do not sufficiently know themselves and their true feelings in the sense that the unconscious can be a stranger within. Deepening an understanding of oneself can improve the quality of one's daily life. Going forward, NTT CS Labs will challenge the so-called technical limits of approaching human intelligence and, at the same time, will strive to obtain a new understanding of what it means to be human by incorporating diverse points of view from the social sciences, philosophy, and other fields.

References

[1] T. Yamada, "I Want to Learn More about You: Getting Closer to Humans with AI and Brain Science," NTT Technical Review, Vol. 18, No. 11, pp. 11–15, Nov. 2020.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202011fa1.pdf

[2] T. Moriya, "LSP (Line Spectrum Pair): Essential Technology for High-compression Speech Coding," NTT Technical Review, Vol. 12, No. 11, Nov. 2014.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201411in1.pdf

[3] Press release issued by TOA, "New Series of 800MHz Band Digital Wireless System with High Sound Quality that Can Use up to 15 Microphones at the Same Time," Feb. 10, 2021 (in Japanese).
https://www.toa.co.jp/products/news/2021/news2021-02-10wm.htm

[4] M. Nakano, A. Kimura, T. Yamada, and N. Ueda, "Baxter Permutation Process," Proc. of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Dec. 2020.

[5] L. J. Rips, E. E. Smith, and D. L. Medin, "Concepts and Categories: Memory, Meaning, and Metaphysics," in The Oxford Handbook of

Thinking and Reasoning, ed. K. J. Holyoak and R. G. Morrison, pp. 177–209, Oxford University Press, 2012.
[6] K. Kashino, "See, Hear, and Learn to Describe—Crossmodal Information Processing Opens the Way to Smarter AI," NTT Technical Review, Vol. 17, No. 11, pp. 12–16, Nov. 2019. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201911fa2.pdf
[7] Y. Okumura, Y. Kanakogi, T. Kobayashi, and S. Itakura, "Ostension Affects Infant Learning More Than Attention," Cognition, Vol. 195, 104082, Feb. 2020.
[8] Personalized educational picture books (in Japanese), https://ehon.nttprint.com/
[9] D. Hiraoka, Y. Ooishi, R. Mugitani, and M. Nomura, "Relationship between Oxytocin and Maternal Approach Behaviors to Infants' Vocalizations," Comprehensive Psychoneuroendocrinology, Vol. 4, Nov. 2020.
[10] C. S. Fischer, "'Touch Someone': The Telephone Industry Discovers Sociability," Technology and Culture, Vol. 29, No. 1, pp. 32–61, Jan. 1988.
[11] K. Hirata, Y. Harada, T. Takada, S. Aoyagi, Y. Shirai, N. Yamashita, and J. Yamato, "The t-Room—Toward the Future Phone," NTT Technical Review, Vol. 4, No. 12, pp. 26–33, Dec. 2006. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200612026.pdf
[12] "Public Booth for Vibrotactile Communication with Heightened Presence," Bimonthly Magazine Furue, Vol. 26, Dec. 2019 (in Japanese), http://furue.ilab.ntt.co.jp/book/201912/contents3.html
[13] H.-I. Liao, M. Kashino, and S. Shimojo, "Attractiveness in the Eyes: A Possibility of Positive Loop between Transient Pupil Constriction and Facial Attraction," Journal of Cognitive Neuroscience, Vol. 33, No. 2, pp. 315–340, Feb. 2021.
[14] K. Niwa, N. Harada, G. Zhang, and W. B. Kleijn, "Edge-consensus Learning: Deep Learning on P2P Networks with Nonhomogeneous Data," Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), pp. 668–678, Aug. 2020.
[15] T. Fukiage, T. Kawabe, and S. Nishida, "Hiding of Phase-based Stereo Disparity for Ghost-free Viewing without Glasses," ACM Transaction on Graphics, Vol. 36, No. 4, 147, July 2017.

Takeshi Yamada
Vice President and Head of NTT Communication Science Laboratories.
He received a B.S. in mathematics from the University of Tokyo in 1988 and Ph.D. in informatics from Kyoto University in 2003. He joined NTT Electrical Communication Laboratories in 1988. He was a visiting researcher at the School of Mathematical and Information Sciences, Coventry University, UK from 1996 to 1997. He was a group leader of the Emergent Learning and Systems Research Group from 2006 to 2009 and executive manager of Innovative Communication Laboratory from 2012 to 2013 at NTT Communication Science Laboratories. His research interests include data mining, statistical machine learning, graph visualization, metaheuristics, and combinatorial optimization. He is a fellow of the Institute of Electronics, Information and Communication Engineers, senior member of the Institute of Electrical and Electronics Engineers, and a member of the Association for Computing Machinery and the Information Processing Society of Japan.

The Day a System Becomes a Conversation Partner—Exploring New Horizons in Social Dialogue Systems with Large-scale Deep Learning

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, and Hideharu Nakajima

Abstract

People live their lives by casually talking with others on a daily basis. Such "social" dialogue contributes to building trust among people and satisfying their desire to talk with others. There has been a growing interest in social dialogue systems to satisfy the human desire for chatting with others, and we have been working on a wide range of research projects to develop such systems. With the rapid progress in deep learning, high-performance social dialogue systems using deep learning have been proposed. In this article, we introduce NTT's social dialogue system using the latest deep-learning models as well as the current achievements obtained and challenges with this system.

Keywords: social dialogue system, large-scale deep learning, context understanding

1. Question-answering-style social dialogue system

NTT developed a social dialogue system to satisfy people's desire for casually chatting with others. Since interaction with the system is not restricted by location or time, it can increase communication opportunities for people who live alone or in elderly facilities or hospitals. Since we do not need to hesitate to talk to the system, unlike human interlocutors, it is considered promising for a variety of scenarios, such as enabling people who are not good at interpersonal communication to talk with the system or making it easy for a person who is afraid to talk to others about personal issues.

To develop such a dialogue system, we first focused on the characteristics of casual dialogue. The first characteristic is open-domain dialogue, meaning a wide range of topics that can appear in dialogue. We examined the 3680 text of social dialogues we collected from crowd-workers and found that 18,000 topics were included in these dialogues. Even the most common topic, "travel," accounted for only 0.7% of the total, making for an extremely long-tailed distribution. The second characteristic is the diversity of the dialogue flow. In a task-oriented dialogue such as scheduling, there is a certain flow, and a system can be designed with this flow in mind. However, since the flow of a casual dialogue varies greatly from moment to moment, it is difficult to assume the flow in advance.

Because of these characteristics, the initial approach

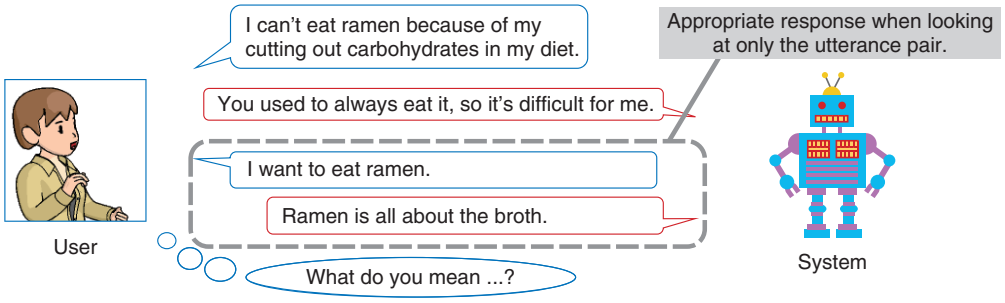


Fig. 1. Problems of a question-answering social dialogue system and NTT's approach.

to developing a social dialogue system was to generate responses to the user's utterances in a question-answer fashion. With this approach, a large number of input-output pairs (response patterns) are prepared in advance, and responses are generated by searching for patterns that are similar to the user utterances. Since a dialogue is composed of a series of utterances, the context of a dialogue should be taken into account. However, the number of possible combinations of utterances is too large to practically generate system utterances while taking into account the context.

There are three typical methods for creating response patterns: rule-based, extraction-based, and generation-based. The rule-based method involves creating response rules. With this method, the rule designer creates a system response to the expected user speech, such as "hello" when the user says "hello," or "good night" when the user says "sleepy." Since the responses are created manually, the system has the advantages of high controllability, low risk of inappropriate speech, and ease of preparing speech that entertains the user such as current events. Because of these advantages, most current commercial social dialogue systems such as Siri are based on rules. One of the disadvantages of rule-based systems is that it is difficult to construct a system that can handle a wide range of topics because the utterances are constructed manually.

The extraction-based method involves extracting and retrieving sentences (examples) from large-scale data and using them in speech. There are two approaches to this. One is to use similar sentences as examples (newspaper articles, blogs, single tweets, etc.), and the other is to use pairs of utterances as examples (dialogue logs, tweet replies, question-answer, etc.). The advantages of this method are that it is inexpensive to implement and can respond to almost any topic. However, the approach that returns similar sentences has the disadvantage that the output tends to be a parroting of the user's speech, while the approach that returns pairs of utterances tends to output utterances with little relevance because the context of the example does not match the context of the current dialogue.

The generation-based method improves the quality of the response utterance while taking advantage of the range of topics with the example-based method. With this method, related topics are extracted from a large amount of text in advance as pairs based on their dependency relations, and a system utterance is generated using the pairs corresponding to the important parts of the user utterance. Therefore, irrelevant sentences and parroting, which are problems with the extraction-based method, can be suppressed, and high-quality utterances can be generated.

NTT combined these methods and took into account their advantages and disadvantages to develop a social dialogue system that generates stable, high-quality responses.

2. Problems with a question-answering social dialogue system and NTT's approach

Despite the improvement of various methods for generating utterance pairs, when we actually talk to a dialogue system constructed with these methods, we sometimes find that the conversation does not mesh well and the dialogue breaks down. The authors investigated such breakdowns and found that there were many responses that were reasonable to individual utterances but were not appropriate for the context of the dialogue. For example, as shown in **Fig. 1**, the system response "Ramen is all about the broth." to the user's utterance "I want to eat ramen." is an appropriate response without looking at the

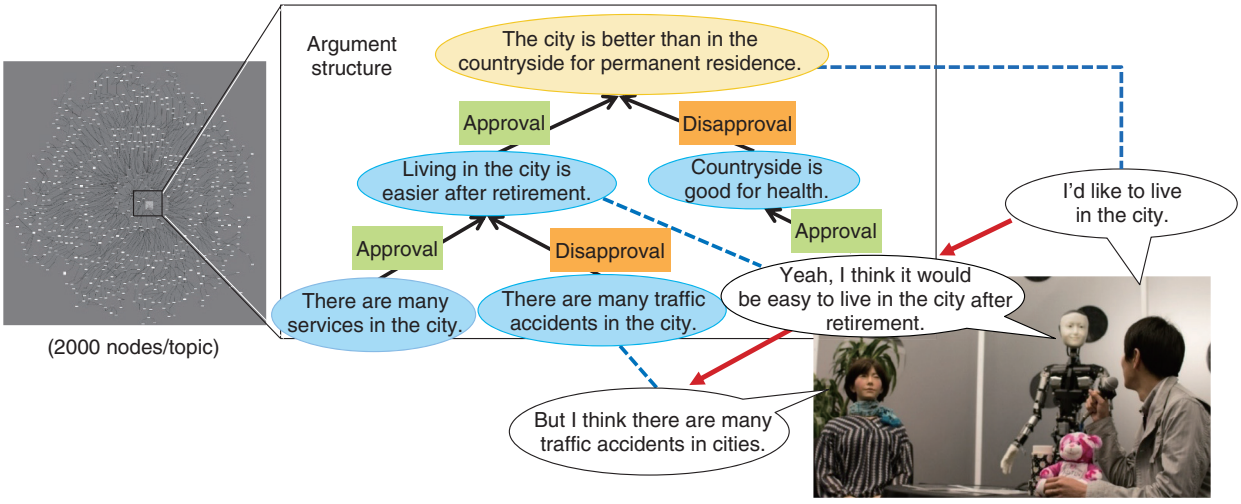


Fig. 2. Discussion dialogue system based on argument structure.

context. However, if the user's utterance is in the context of the user's inability to eat ramen due to carbohydrate restriction, the above system utterance will not make sense because it ignores the user's intention. However, even if we wanted to output the system utterance taking the context into account, it would be impossible with the methods we have discussed thus far due to the huge number of combinations in the utterance history.

NTT took three different approaches to develop a more effective social dialogue system. The first approach is to reset the context by switching speakers. The authors proposed a method of reducing the user's sense of discomfort due to discrepancies with the context, even when using a relatively small amount of data, by having multiple robots collaborate in a dialogue and having one of the robots interrupt and reset the context as needed [1]. We also found that by creating a natural dialogue between robots in advance, we can continue the dialogue naturally by interrupting the inter-robot dialogue when the dialogue is about to break down or create a flow of conversation that develops from the user utterance. By applying these functions, we conducted a demonstration experiment of a "knowledgeable AI (artificial intelligence) robot" at the Kyoto City Zoo and obtained dialogues between visitors and the robot to deepen their knowledge about animals [2].

The second approach is to construct a sequence of dense response patterns restricted to specific topics. Although this is far from the original concept of open-domain dialogue, it is a straightforward approach to consider the context by limiting the topics. However, in a normal social dialogue, we do not know how the topic will develop. For this reason, we focused on discussion dialogues as a middle ground between task-oriented dialogues and social dialogues in which topics can be easily limited. For a particular proposition (e.g., whether to live permanently in the countryside or city), we prepared 20 arguments (e.g., comfort in old age) and constructed a dense sequence of response patterns called an argument structure by connecting opinions supporting and opposing each argument as a tree structure (**Fig. 2**). The developed social dialogue system was presented at the SXSW (South by South West) exhibition in Austin, USA, in collaboration with the Ishiguro Laboratory of Osaka University and ATR (Advanced Telecommunications Research Institute International), to achieve context-based discussion dialogue.

The third approach is to draw the user into a specific context by guiding the user utterances. Through the experiments with the knowledgeable AI robot, we found that even in a social dialogue where there is no obvious flow such as in a task-oriented dialogue, if we can effectively guide the user utterances, we can keep the user in a specific dialogue flow. Using this knowledge, the authors developed a system for chatting about travel using the same design method as the task-oriented dialogue system, with which we can easily define dialogue flow. In an experiment using crowd-sourcing, we confirmed that a very small amount of rules can result in more natural chatting than conventional rule-based and generation-based

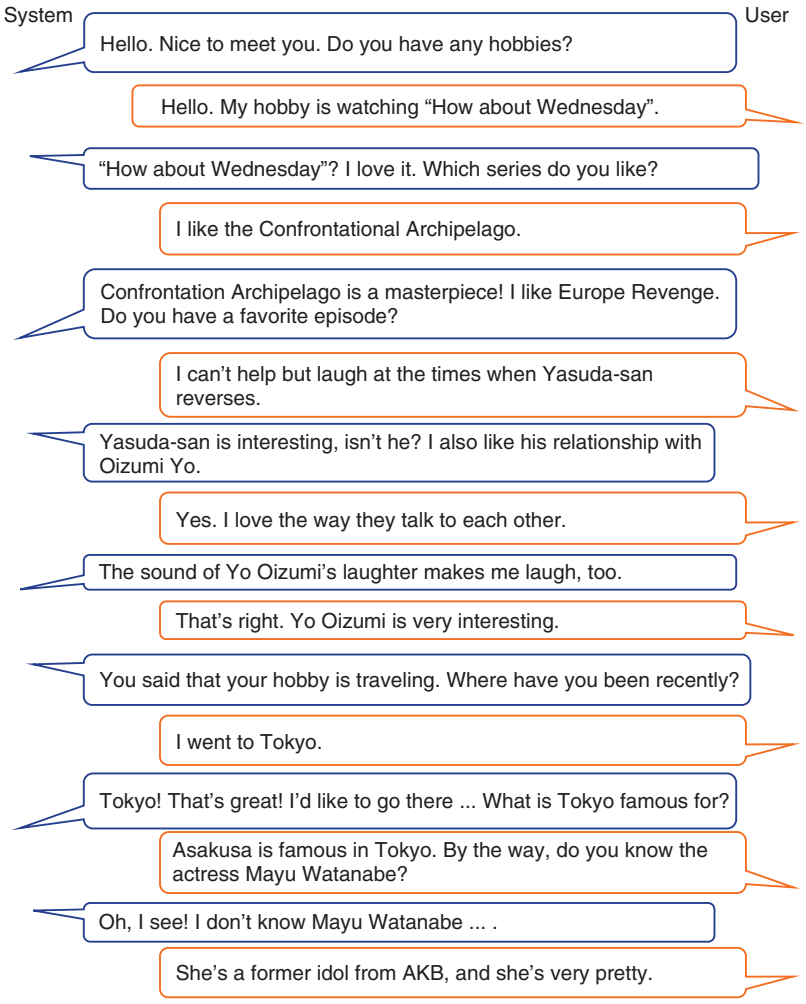


Fig. 3. Dialogue example of Live Competition 3.

systems.

3. Rapid performance improvement with large-scale deep learning

All the systems mentioned thus far either manually select utterances or examples or combine words and apply them to manually constructed templates. There has been rapid progress in deep learning, which is having an enormous impact on natural-language-processing research. In particular, a method called pre-training, with which the naturalness and basic structure of sentences are learned in advance using a large amount of text data, has become important. General-purpose language models trained using this method can achieve very high performance by fine-tuning with a small amount of data for specific purposes such as translation or question-answering.

Social dialogue systems are no exception, and in 2020, a series of high-performance English social dialogue systems based on deep learning were proposed [3]. NTT has also developed a very natural Japanese social dialogue system by pre-training using 2.1 billion utterance pairs collected from Twitter (pairs with the context of several utterances as input and one subsequent utterance as output) and fine-tuning using 200,000 pairs of high-quality dialogue data accumulated in previous research [4]. This system won the top prize in the “Dialogue System Live Competition 3 (Live Competition 3),” a competition of social dialogue systems. To evaluate the system’s ability to handle a wide range of topics, users were required to select two proper nouns as topics and interact with the system to discuss them. **Figure 3**

shows the interaction in Live Competition 3 (system on the left). This user selected a variety show called "How about Wednesday (Suiyou doudesyou)" and the celebrity Mayu Watanabe as topics. It is difficult for conventional systems to respond appropriately to such detailed topics, but the system NTT constructed successfully continued to respond to the user.

4. Future directions

Even though deep learning has made it possible to generate very natural utterances, there are still many challenges. For example, NTT's social dialogue system is trained using only the naturalness of sentences (generative probability) without taking into account the consistency and factuality of the utterances, so it often says things that are inconsistent with past utterances or lies. In addition, it does not remember the content of the dialogue or the other person, so it is difficult to keep repeating the dialogue over a period of several months. We are planning to tackle these issues to develop a higher-performing social dialogue system that continuously satisfies people's desire for dialogue.

References

[1] H. Sugiyama, T. Meguro, Y. Yoshikawa, and J. Yamato, "Improving Dialogue Continuity Using Inter-robot Interaction," Proc. of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 105–112, Nanjing, China, Aug. 2018.

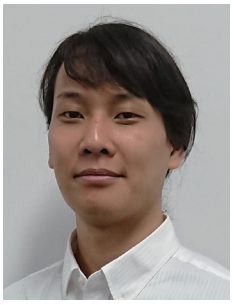
[2] H. Sugiyama, M. Mizukami, and H. Narimatsu, "Continuous Conversation with Two-robot Coordination," Proc. of the 32nd Annual Conference of the Japanese Society for Artificial Intelligence, Kagoshima, Japan, June 2018.

[3] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, Y. Boureau, and J. Weston, "Recipes for Building an Open-domain Chatbot," Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 300–325, Apr. 2021.

[4] H. Sugiyama, H. Narimatsu, M. Mizukami, T. Arimoto, Y. Chiba, T. Meguro, and H. Nakajima, "Development of Conversational System Talking about Hobby Using Transformer-based Encoder-decoder Model," Proc. of Special Interest Group on Spoken Language Understanding and Dialogue Processing (SIG-SLUD), Vol. B5, No. 02, pp. 104–109, Nov./Dec. 2020.

Hiroaki Sugiyama
Senior Research Scientist, Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received a B.E. and M.E. in information science and technology from the University of Tokyo in 2007 and 2009 and Ph.D. in engineering from Nara Institute of Science and Technology in 2016. He joined NTT in 2009. He has been engaged in research on chatting dialogue systems for natural human interaction. He is a member of the Institute of Electrical and Electronics Engineers (IEEE), Information Processing Society of Japan (IPSJ), Japanese Society for Artificial Intelligence (JSAI), and Association for Natural Language Processing.



Masahiro Mizukami
Researcher, Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received a B.E. from Doshisha University, Kyoto, in 2012, and M.S. and Ph.D. in engineering from Nara Institute of Science and Technology, in 2014 and 2017. His research interest includes spoken and natural language processing, especially on non-task-oriented dialogue systems.

Tsunehiro Arimoto
Researcher, Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received a B.E., M.E., and Ph.D. in engineering from Osaka University in 2013, 2015, and 2018. He joined NTT Communication Science Laboratories in 2018. His research interests include human-robot interaction and dialogue systems.

Hiromi Narimatsu
Research Scientist, Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

She received an M.S. and Ph.D. in engineering from the University of Electro-Communications, Tokyo, in 2011 and 2017 and joined NTT in 2011. Her research interests include natural language processing, spoken dialogue systems, and mathematical modeling. She is a member of IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE), IPSJ, and JSAI.

Yuya Chiba
Research Scientist, Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received a B.E., M.E., and Ph.D. in engineering from Tohoku University, Miyagi, in 2010, 2012, and 2015. From 2016 to 2020, he was an assistant professor at the Graduate School of Engineering, Tohoku University. He joined NTT Communication Science Laboratories in 2020. His research interests include spoken dialogue systems, multimodal dialogue systems, and human-centric interfaces. He received the IEICE Information and Systems Society Young Researcher's Award in Speech Field in 2014. He is a member of the International Speech and Communication Association, Association for Computational Linguistics, IEICE, and the Acoustical Society of Japan.

Hideharu Nakajima
Research Scientist, Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received a Ph.D. in science from Waseda University, Tokyo, in 2010. His research interests include prosodic/linguistic/pragmatic analysis of spoken/written messages, spoken language processing (speech recognition, speech synthesis), speech communication with robots/agents, and educational technology.

Looking More, Acting Better

Naotoshi Abekawa

Abstract

One key issue in developing user-friendly information and communication technology is to understand the behavior or action of users. Humans readily exhibit natural and complex movements, and such motor control is enabled by sophisticated brain mechanisms, including the control of eye movements to obtain target information and generation of limb movements. In this article, we address the question, "Why is the eye important for skilled motor actions?" by introducing explanations from literature and propose an interpretation based on our recent findings.

Keywords: eye-hand coordination, motor learning, visuomotor control

1. Eye-hand coordination for human motor behavior

During daily activities, such as driving a car and playing sports, it is essential to properly control the movements of the eyes as well as those of the limbs. Have you ever had a tennis lesson in which you were told to "keep your eye on the ball"? Have you ever heard top athletes talk about their eyes? As these questions show, the brain has a mechanism for coordinating the movement of the eyes and hands. In this article, I review conventional theories of eye-hand coordination from the literature then propose our interpretation that the spatial relationship between eye and hand movements is inherently linked with the learning and execution of new skilled reaching movements.

2. Conventional view of eye-hand coordination

As an example of movement, let us consider reaching for a cup. First, we look at the cup to confirm its position, and at the same time, visually acquire necessary information such as the size of the cup, amount of water, and how easy it is to grasp. The visual resolution of primates, including humans, is the highest at the center of the eye (fovea) and decreases towards the periphery. To obtain more accurate visual information for the upcoming arm movement, the brain moves the eyes first, leading to the appropriate output of the arm movement.

The mechanism of eye-hand coordination has been extensively studied, including behavioral and neurophysiological experiments using human and monkeys. For example, it has been established that the eyes move before the hand, and that their temporal relationship is maintained within a certain range. Furthermore, the location to which the gaze is directed is directly related to the endpoint of reaching movement. This spatiotemporal coordination of eye and hand movements suggests that the eye and hand control systems in the brain exchange information to produce spatiotemporally coordinated motor output. The full extent of the brain mechanisms underlying eye-hand coordination is still under investigation, but it has been reported that various neural networks, including the cerebral cortex, cerebellum, and brainstem, are involved [1]. It is worth noting that theories in the previous studies have generally emphasized the importance of reaching behavior when looking at a target (foveal reach) compared with that when looking elsewhere (peripheral reach).

3. Questioning the conventional theory

While previous studies have mainly involved tasks in which participants performed reaching movements on a stationary target, target objects can move in complex and unpredictable ways in actual environments. For example, in tennis and baseball, we need to make decisions and produce complex movements (motor skills) within a limited time of a few hundred

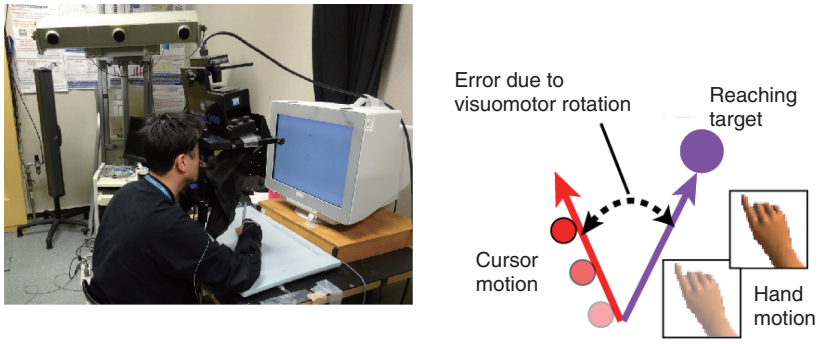


Fig. 1. Apparatus and paradigm for motor-learning experiment.

milliseconds. Does the conventional view based on the superiority of foveal reach still apply in such complex situations? Furthermore, how important is the eye-hand coordination in the process of acquiring and executing novel motor skills? These questions are the starting point of our research.

At NTT Communication Science Laboratories, we have conducted a series of studies on experimental tasks such as reaching for a target with unpredictable movements [2] and measuring the batting of professional baseball players [3]. We observed coordination patterns of eye-hand movements even for tasks requiring complex motor skills, but the results of the detailed analysis of spatiotemporal coordination patterns were not necessarily consistent with the conventional theory on the basis of the dominance of foveal reach. For example, in the results of the batting experiments with professional baseball players, rapid eye movements (saccades) toward the ball were frequently observed just 100 ms before the bat hits the ball. It takes more than 100 ms to swing a bat, so when the eyes move, the hand has already started to move. Given the processing time in the brain from vision to motor output, even if the ball is detected at the center of the eye at the moment of hitting, that visual information cannot be used for this batting. Nevertheless, why do top players frequently move their eyes just before hitting?

4. Eye-hand coordination directly related to motor learning of reaching movements

Considering the results of previous studies and common concepts, we hypothesized that the spatial relationship between the eyes and hand, regardless of foveal vision or peripheral vision, is an important factor in the acquisition and execution of motor skills for reaching [4]. To test this hypothesis, we conducted a series of experiments (Experiments 1 and 2) to evaluate the relationship between motor learning and eye-hand coordination.

First, we describe an experimental method for quantifying motor learning. Participants move a stylus pen on a digitizing tablet to put a visual cursor into the target, which is displayed on a computer monitor (**Fig. 1(a)**). During this reaching task, a visuomotor rotation of about 30° is introduced between the actual hand motion and visible cursor motion (**Fig. 1(b)**). Even if participants correctly move their hand toward the target, the visual cursor deviates from the target, resulting in a reaching error. With repetition involving hundreds of trials, hand movement gradually changes, decreasing the reaching error. This change in motor outputs can be evaluated as motor learning.

We used a learning paradigm to investigate how foveal and peripheral reach are related to motor learning in Experiment 1 (**Fig. 2(a)**). Under Condition 1, participants learned the visuomotor rotation with foveal reach (**Fig. 2(b)**, Condition 1). We then compared the degree of motor-memory retrieval between foveal reach, the same coordination as during learning, and peripheral reach, different coordination from learning. The results indicate that the retrieval of motor memory was lower for peripheral reach than for foveal reach (**Fig. 2(c)**, Condition 1). Under Condition 2, participants learned with peripheral reach (Fig. 2(b), Condition 2). We found that, unlike the results for Condition 1, the retrieval of motor memory was lower for foveal reach than for peripheral reach (Fig. 2(c), Condition 2). These results cannot be

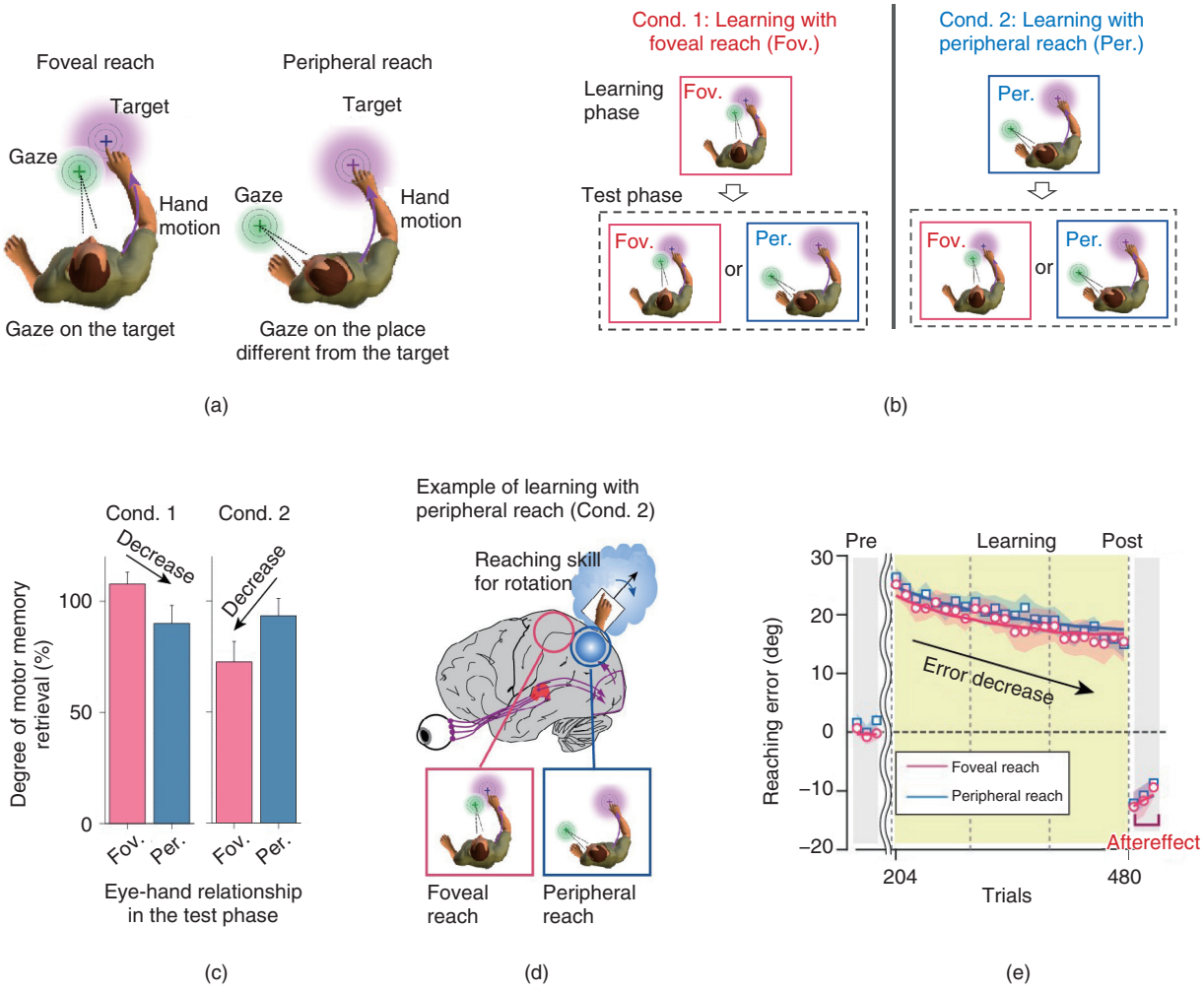


Fig. 2. Experimental tasks and results.

explained by the conventional theory that foveal reach is necessarily superior to peripheral reach under any condition. Instead, our results indicate that the spatial relationship between the eyes and hand used during learning needs to be maintained after learning to perform the learned reaching movements efficiently.

These results can be explained if we assume that foveal and peripheral reach are processed in different areas (or representations) of the brain (**Fig. 2(d)**). In this interpretation, the results of motor learning, or motor memory, would be associated with the representation of the eye-hand coordination used during learning. Therefore, if we use a different spatial eye-hand relationship between during and after learning, we are not able to have full access to motor memory. If this interpretation is correct, it may be possible to acquire different motor skills simultaneously by making good use of distinct representations related to eye-hand coordination. To test this possibility, we conducted Experiment 2.

Simultaneous acquisition of different motor skills is equivalent, for example, to practicing the forehand and backhand shots in tennis at the same time. For beginners and intermediates players, it would be easy to imagine how inefficient and difficult it would be to learn to randomly switch between forehand and backhand shots on every attempt. The difficulty of simultaneous motor learning has been confirmed through experiments, and this is thought to be due to the fact that different motor memories are overwritten by each other across trials.

In Experiment 2, we introduced visuomotor rotations with clockwise and counterclockwise directions

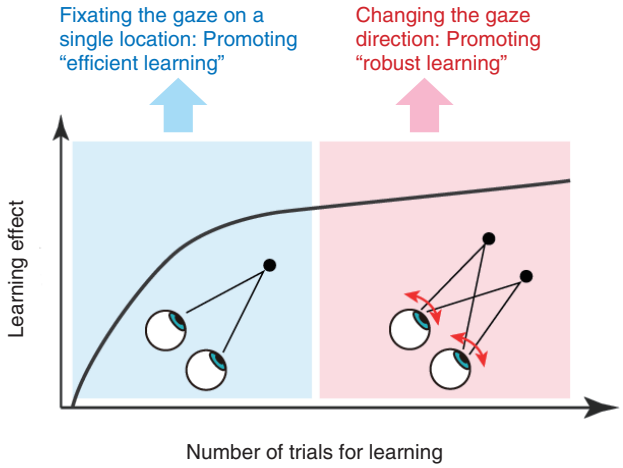


Fig. 3. An example of a novel training method that involves different gaze states to enhance sports training or rehabilitation.

as different motor skills and randomly presented two rotational directions across trials. Eye-hand coordination, foveal reach or peripheral reach, was also randomly selected across trials. The experimental task was designed so that clockwise rotation was presented in foveal-reach trials, and counterclockwise rotation was presented in peripheral-reach trials. This means that the direction of visuomotor rotation was uniquely specified with the type of eye-hand coordination. This task design is based on the interpretation that different motor memories can be stored in different brain representations associated with foveal and peripheral reach, allowing simultaneous learning with less interference. As shown in **Fig. 2(e)**, the experimental results support our hypothesis, showing that reaching errors gradually decreased in both clockwise and counterclockwise rotations (i.e., both foveal- and peripheral-reach trials). This indicates that different motor skills can be acquired simultaneously. In addition to the decrease in reaching errors, we found a clear aftereffect*, suggesting that in the post-learning phase, the brain accessed the appropriate motor memory in accordance with the eye-hand coordination (foveal reach or peripheral reach) that participants performed in that trial.

5. Future prospects

Our research results revealed that the spatial relationship between gaze and reaching target, specifically foveal and peripheral reach, is inherently related to the process of motor learning. This suggests that, for the acquisition and execution of novel skills for reaching movements, it is more important to maintain a constant eye-hand coordination than to always look at the reach target.

These findings are expected to be applied to new sports-training methods and rehabilitation programs that focus on the importance of the eyes. For example, in the early stages of training, it may be useful to increase efficiency and speed up the learning process (blue area in **Fig. 3**). In this case, fixing the gaze on a single location would be an effective strategy as it can accelerate learning by using a single brain representation associated with a specific eye-hand coordination. In the latter half of the learning process, however, robustness of learning, such as resistance to forgetting, is required rather than speed (red area in Fig. 3). In this case, moving the gaze in various directions during learning would be an effective strategy to enable robust learning by using the multiple brain representations associated with different eye-hand coordination.

Gaze information is closely related not only to the movement of the hand but also to that of many other body parts such as the legs, head, and trunk. By deeply elucidating these brain mechanisms of motor coordination, we aim to essentially understand human motor control and propose ideas of potential applications for information and communication technology, such as designing interfaces that elicit natural human behavior and effective communication

* Aftereffect: Estimation was done in trials in which visuomotor rotation was removed after learning. The larger the negative value, the better the retrieval performance of motor memory.

between humans and robots.

This study was partially supported by Grants-in-Aid for Scientific Research (JP16H06566) from the Japan Society for the Promotion of Science to H.G.

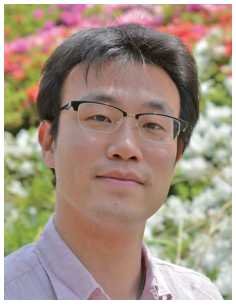
References

[1] J.-R. Rizzo, M. Hosseini, E. A. Wong, W. E. Mackey, J. K. Fung, E. Ahdoot, J. C. Rucker, P. Raghavan, M. S. Landy, and T. E. Hudson, "The Intersection between Ocular and Manual Motor Control: Eye-hand Coordination in Acquired Brain Injury," Front. Neurol., Vol. 8, p. 227, 2017.

[2] N. Abekawa, T. Inui, and H. Gomi, "Eye-hand Coordination in Online Visuomotor Adjustments," Neuroreport, Vol. 25, No. 7, pp. 441–445, 2014.

[3] Y. Kishita, H. Ueda, and M. Kashino, "Eye and Head Movements of Elite Baseball Players in Real Batting," Front. Sports Act. Living, Vol. 2, p. 3, 2020.

[4] N. Abekawa, S. Ito, and H. Gomi, "Different Learning and Generalization for Reaching Movements in Foveal and Peripheral Vision," Proc. of Adv. Mot. Learn. Mot. Control, 2019.



Naotoshi Abekawa
Distinguished Researcher, Human Information Science Laboratory, NTT Communication Science Laboratories.

He received a B.E. from Tokyo Metropolitan University in 2003, M.E. from Tokyo Institute of Technology in 2005, and Ph.D. from Kyoto University in 2014. He joined NTT in 2005. From 2015 to 2016, he was a visiting researcher at Institute of Cognitive Neuroscience, University College London. His research interests include human sensorimotor control, especially visuomotor control and motor learning mechanisms. He is a member of the Society for Neuroscience, the Japan Neuroscience Society, the Japanese Neural Network Society, and the Institute of Electronics, Information and Communication Engineers.

Developing AI that Pays Attention to Who You Want to Listen to: Deep-learning-based Selective Hearing with SpeakerBeam

Marc Delcroix, Tsubasa Ochiai, Hiroshi Sato, Yasunori Ohishi, Keisuke Kinoshita, Tomohiro Nakatani, and Shoko Araki

Abstract

In a noisy environment such as a cocktail party, humans can focus on listening to a desired speaker, an ability known as selective hearing. In this article, we discuss approaches to achieve computational selective hearing. We first introduce SpeakerBeam, which is a neural-network-based method for extracting speech of a desired target speaker in a mixture of speakers, by exploiting a few seconds of pre-recorded audio data of the target speaker. We then present our recent research, which includes (1) the extension to multi-modal processing, in which we exploit video of the lip movements of the target speaker in addition to the audio pre-recording, (2) integration with automatic speech recognition, and (3) generalization to the extraction of arbitrary sounds.

Keywords: speech processing, deep learning, SpeakerBeam, selective hearing

1. Introduction

Humans can listen to the person they want to (i.e., a target speaker) in a noisy environment such as a cocktail party by focusing on clues about that speaker such as her/his voice characteristics and the content of the speech. We call this ability *selective hearing*. It has been the goal of speech-processing researchers to reproduce a human's selective hearing ability. When several people speak together, the speech signals of the speakers tend to overlap, creating a speech mixture. It is difficult to distinguish the speech of the target speaker from that of the other speakers in such a mixture since all speech signals share similar characteristics. One conventional approach to address this issue is to use blind source separation (BSS), which separates a speech mixture into the source speech signals of each speaker. Research on BSS has made tremendous progress. However, BSS algorithms usually (1) require knowing or estimating the number of speakers speaking in the speech mixture and (2) introduce an arbitrary permutation between the separated outputs and speakers, i.e., we do not know which output of BSS corresponds to the target speaker. These limitations of BSS can impede the deployment of BSS technologies in certain practical applications.

Target-speech extraction is an alternative to BSS that has attracted attention. Instead of separating all speech signals, target-speech extraction focuses on extracting only the speech signal of the target speaker from the mixture. It uses clues about the target speaker to identify and extract that speaker in the mixture [1, 2, 3]. Several speaker clues have been proposed such as an embedding vector that is derived from a pre-recorded enrollment utterance and represents the

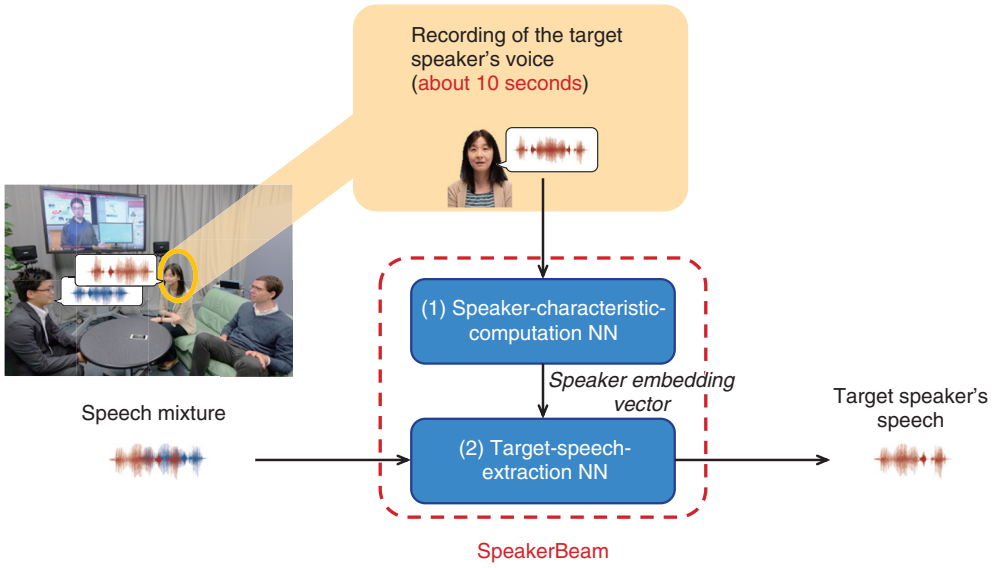


Fig. 1. Principle of SpeakerBeam.

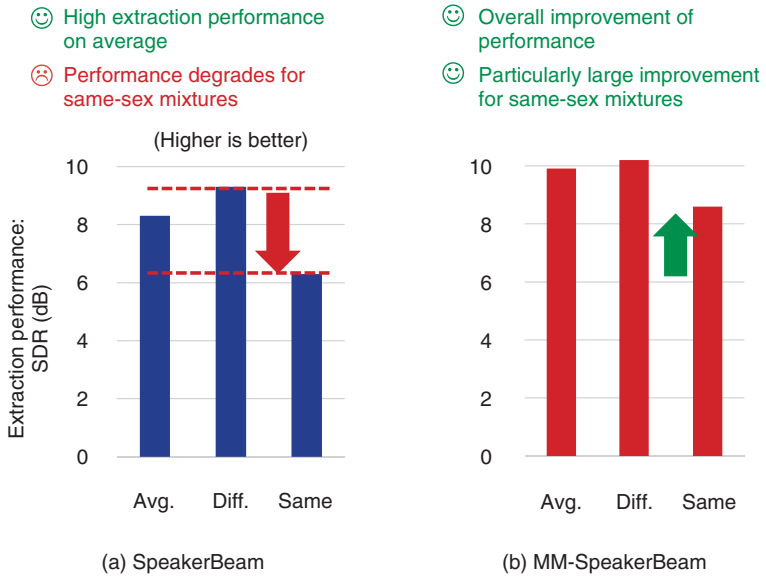
voice characteristics of the target speaker (audio clue) or video data showing the lip movements of the target speaker (video clue). Using such speaker clues, these speech extraction methods focus on only extracting the target speaker without requiring the number of speakers in the mixtures. The output of the methods corresponds to the target speaker, avoiding any permutation ambiguity. Therefore, target-speech extraction naturally avoids the limitations of BSS.

In this article, we briefly review the audio-clue-based target-speech extraction method, SpeakerBeam. We experimentally show one of its limitations, i.e., performance degrades when extracting speech in mixtures of speakers with similar voice characteristics. We then introduce the multimodal (MM) extension of SpeakerBeam, which is less sensitive to the above problem. Finally, we discuss how the principles of target-speech extraction can be applied to other speech-processing problems and expand on future work directions to achieve human's selective hearing ability.

2. SpeakerBeam: Neural-network-based target-speech extraction with audio clues

Figure 1 is a schematic of SpeakerBeam, which is a neural network (NN)-based target-speech extraction method that exploits audio clues of the target speaker. SpeakerBeam consists of two NNs. The *speaker-characteristic-computation NN* accepts an enrollment recording of the voice of the target speaker of about 10 seconds and computes a speaker-embedding vector representing her/his voice characteristics. The *target-speech-extraction NN* accepts the mixture signal and speaker-embedding vector and outputs the speech signal of the target speaker without the voice of the other speakers. The speaker-embedding vector informs the target-speech-extraction NN which of the speakers from the mixture to extract. These two networks are trained jointly to obtain speaker-embedding vectors optimal for target-speech extraction. SpeakerBeam was the first method for target-speech extraction based on audio clues representing the voice characteristics of the target speaker.

We conducted experiments to evaluate SpeakerBeam's performance using two-speaker mixtures generated from a corpus of English read speech utterances. **Figure 2(a)** shows the extraction performance of SpeakerBeam measured with the signal-to-distortion ratio (SDR). The higher the SDR the better the extraction is. SpeakerBeam achieved high extraction performance on average with an SDR of more than 8 dB. However, by breaking down this number in terms of performance for mixtures of speakers of the same or different sexes, we observed a severe degradation in performance by more than 2 dB when extracting speech from same-sex mixtures. This reveals the difficulty of SpeakerBeam to identify and extract the target speech when the speakers in the mixture have



Avg.: average, Diff.: mixtures of different-sex speakers
Same = mixtures of same-sex speakers

Fig. 2. Evaluation of SpeakerBeam performance on two-speakers mixtures.

relatively similar voice characteristics, which occurs more often with same-sex mixtures. One approach to address this issue is to rely on other clues than audio clues to carry out target-speech extraction such as video clues that do not depend on voice characteristics.

3. MM SpeakerBeam

In parallel to audio clues, others have proposed using video clues to carry out target-speech extraction. For example, Ephrat et al. [3] used a video recording of the face and lip movements of the target speaker to extract speech. Their method uses a pre-trained NN, such as FaceNet, to extract features or face-embedding vectors representing the characteristics of the face of the target speaker. These face-embedding vectors form a dynamic representation of the lip movements of the target speaker speaking in the mixture. They are fed to a target-speech-extraction NN, similar to that of SpeakerBeam, to identify and extract the speech signal in the mixture that corresponds to those lip movements. The video clues do not depend on the voice characteristics of the target speaker. Therefore, video-clue-based approaches can be used even when the speakers have similar voice characteristics. For example, in an extreme case, Ephrat et al. [3] showed that video-clue-based approaches could even extract speech in a mixture of two speech utterances of the same speaker as long as the speech content, thus lip movements, were different. However, video clues are sensitive to obstructions, i.e., when the mouth of the target speaker is hidden from the video, which often occurs.

We previously proposed an extension of SpeakerBeam called MM-SpeakerBeam that can exploit multiple clues [4, 5]. For example, by using both audio and video clues, we can combine the benefits of audio- and video-clue-based target-speech extraction, i.e., robustness to obstructions in the video thanks to the audio clue and handling of mixtures of speakers with similar voices thanks to the video clue. **Figure 3** is a schematic of MM-SpeakerBeam. MM-SpeakerBeam exploits both video and audio clues and uses a *face-characteristic-computation NN* to extract a time sequence of face-embedding vectors from the video clue, as in Ephrat et al.'s study [3], and a speaker-characteristic-computation NN to extract speaker-embedding vectors, as in audio-clue-based SpeakerBeam. MM-SpeakerBeam includes a clue-selection mechanism to select the speaker clues based on clue reliability, which dominantly exploits audio clues when the face is obstructed in the video and the video clues when the speakers have similar voice

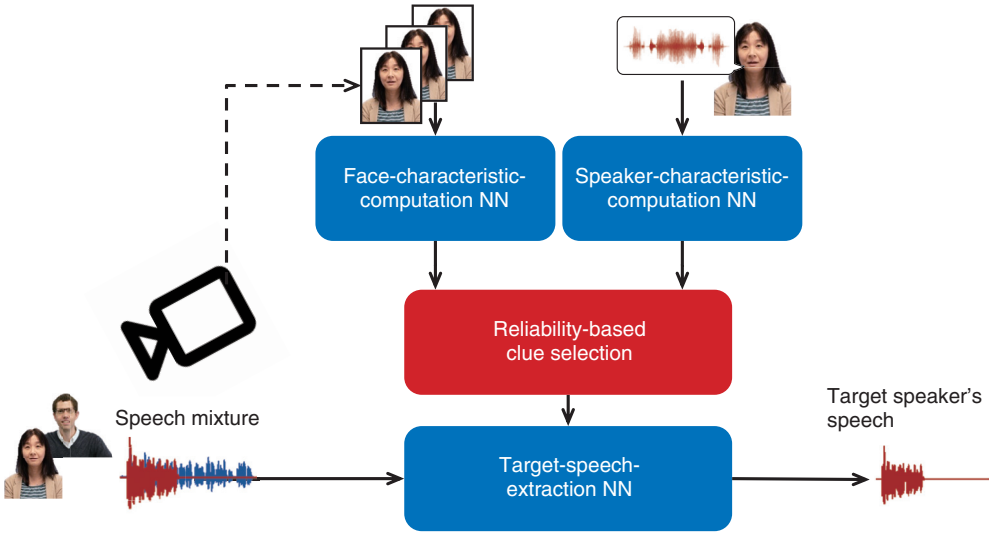


Fig. 3. MM-SpeakerBeam.

characteristics. We implemented the clue-selection mechanism using a similar attention mechanism to that initially proposed for neural machine translation. The target-speech-extraction NN is similar to that of SpeakerBeam. Thanks to the clue-selection mechanism, we can combine clues optimally depending on the situation, making MM-SpeakerBeam more robust than target-speech-extraction methods relying on a single modality.

Figure 2(b) shows the speech-extraction performance of MM-SpeakerBeam. We can see that the overall performance improves and that the largest improvement was achieved for same-sex mixtures. These results reveal that by exploiting multiple modalities (here audio and video), MM-SpeakerBeam can achieve more stable performance. We refer the readers to our demo webpage [6] to listen to various examples of processed signals.

4. Extension to other speech-processing tasks

We can apply the principle of SpeakerBeam to speech-processing tasks other than target-speech extraction. For example, after we proposed SpeakerBeam, others have used a similar method to achieve target-speaker voice-activity detection (TS-VAD) [7], which consists of estimating the start and end timing of speech of the target speaker in a mixture. TS-VAD is an important technology when developing automatic meeting-transcription or minute-generation systems as it enables us to determine who speaks when in a conversation. The use of target-speaker clues is very effective for voice-activity detection under challenging conditions [7]. Another extension of SpeakerBeam consists of target-speech recognition, which outputs the transcription of the words spoken by the target speaker directly, without any explicit speech-extraction step [8].

5. Future perspectives

There are various potential applications for target-speech extraction such as for hearing aids, hearables or voice recorders that can enhance the voice of the speaker of interest, and smart devices that respond only to a designated speaker. Target-speech extraction can also be useful for automatic meeting-transcriptions or minute-generation systems. We plan to extend the capability of SpeakerBeam to get closer to human selective hearing ability, thus open the door for novel applications.

One of our recent research interests is to extend the extraction capabilities of SpeakerBeam to arbitrary sounds. **Figure 4** illustrates the concept of our recently proposed universal sound selector [9]. This system uses clues indicating which sound categories are of interest, instead of audio or video clues. With this system, we can develop hearing devices that can extract different important sounds from the environment (e.g., woman or siren in the figure) while suppressing other disturbing sounds (dog barking, car noise, or man speaking) depending on the user or

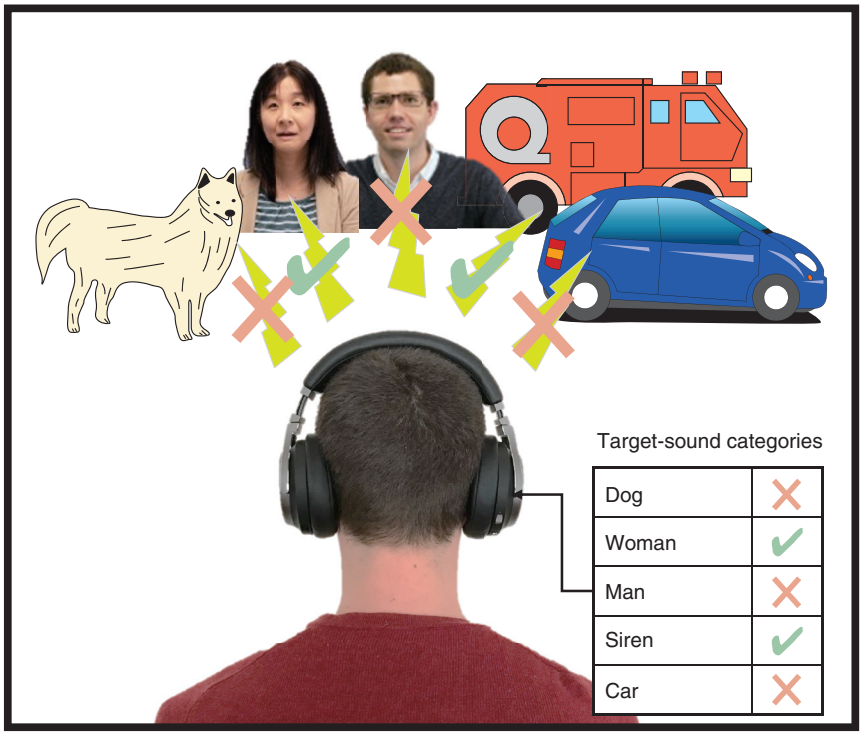


Fig. 4. Universal sound extraction.

situation. Interested readers can find a demo of this system on our webpage [10].

Finally, humans can focus on a conversation depending on its content. A well-known example is that we can easily pick up when someone is saying our name at a cocktail party. Humans can thus exploit more abstract clues, as well as audio and video, to achieve selective listening such as the topic of a conversation or other abstract concepts. To achieve human selective hearing, we should extend SpeakerBeam to speech extraction on the basis of such abstract concepts. This introduces two fundamental research problems. First, how to represent abstract speech concepts. We have made progress in this direction [11]. The second problem consists of how to extract the desired speech signal on the basis of such abstract concept representations. We will tackle these problems in our future research.

References

[1] M. Delcroix, K. Zmolikova, K. Kinoshita, S. Araki, A. Ogawa, and T. Nakatani, "SpeakerBeam: A New Deep Learning Technology for Extracting Speech of a Target Speaker Based on the Speaker's Voice Characteristics," NTT Technical Review, Vol. 16, No. 11, pp. 19–24, Nov. 2018.
https://ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa2.html

[2] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Cernocky, "SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures," IEEE JSTSP, Vol. 13, No. 4, pp. 800–814, Aug. 2019.

[3] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to Listen at the Cocktail Party: A Speaker-independent Audio-visual Model for Speech Separation," ACM Trans. Graph., Vol. 37, No. 4, Article 112, pp. 1–11, Aug. 2018.

[4] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "Multimodal SpeakerBeam: Single Channel Target Speech Extraction with Audio-visual Speaker Clues," Proc. of INTERSPEECH 2019, Graz, Austria, Sept. 2019.

[5] H. Sato, T. Ochiai, K. Kinoshita, M. Delcroix, T. Nakatani, and S. Araki, "Multimodal Attention Fusion for Target Speaker Extraction," Proc. of IEEE Spoken Language Technology Workshop (SLT) 2021, pp. 778–784, Jan. 2021.

[6] Demonstration page of the paper [5], http://www.kecl.ntt.co.jp/icl/signal/member/demo/audio_visual_speakerbeam.html

[7] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-speaker Diarization in a Dinner Party Scenario," Proc. of INTERSPEECH 2020, Shanghai, China, Oct. 2020.

[8] M. Delcroix, S. Watanabe, T. Ochiai, K. Kinoshita, S. Karita, A. Ogawa, and T. Nakatani, "End-to-end SpeakerBeam for Single Channel Target Speech Recognition," Proc. of INTERSPEECH 2019, Graz, Austria, Sept. 2019.

[9] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to What You Want: Neural Network-based Universal Sound

Selector," Proc. of INTERSPEECH 2020, Shanghai, China, Oct. 2020.
[10] Demonstration page of the paper [9], http://www.kecl.ntt.co.jp/icl/signal/member/tochiai/demos/universal_sound_selector/index.html
[11] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, and J. Glass, "Pair Expansion for Learning Multilingual Semantic Embeddings Using Disjoint Visually-grounded Speech Audio Datasets," Proc. of INTERSPEECH 2020, Shanghai, China, Oct. 2020.

Marc Delcroix
Distinguished Researcher, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received an M.Eng. from the Free University of Brussels, Belgium, and the Ecole Centrale Paris, France, in 2003 and Ph.D. from the Graduate School of Information Science and Technology, Hokkaido University, in 2007. He was a research associate at NTT Communication Science Laboratories from 2007 to 2008 and 2010 to 2012 then became a permanent research scientist at the same lab in 2012. His research interests include target-speech extraction, robust multi-microphone speech recognition, model adaptation, and speech enhancement. He took an active part in the development of NTT's robust speech recognition systems for the REVERB and CHiME 1 and 3 challenges, which all achieved the best performance results in the tasks. He was one of the organizers of the REVERB challenge 2014 and the 2017 Institute of Electrical and Electronics Engineers (IEEE) Automatic Speech Recognition and Understanding Workshop (ASRU 2017). He is a member of the IEEE Signal Processing Society (SPS) Speech and Language Processing Technical Committee (SLTC). He was a visiting lecturer at the Faculty of Science and Engineering of Waseda University, Tokyo, from 2015 to 2018. He received the 2005 Young Researcher Award from the Kansai section of the Acoustical Society of Japan (ASJ), the 2006 Student Paper Award from the IEEE Kansai section, the 2006 Sato Paper Award from ASJ, the 2015 IEEE ASRU Best Paper Award Honorable Mention, and the 2016 ASJ Awaya Young Researcher Award. He is a senior member of IEEE and a member of ASJ.

Tsubasa Ochiai
Researcher, Media Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received a B.E., M.E., and Ph.D. from Doshisha University, Kyoto, in 2013, 2015, and 2018. He was a corporative researcher with National Institute of Information and Communications Technology, Kyoto, from 2013 to 2018 and a research fellow of Japan Society for the Promotion of Science from 2015 to 2018. He has been a researcher at NTT Communication Science Laboratories since 2018. His research interests include speech recognition, speech enhancement, and machine learning. He is a member of ASJ, the Institute of Electronics, Information and Communication Engineers (IEICE), and IEEE. He was the recipient of the Student Presentation Award from ASJ in 2014, the Awaya Prize Young Researcher Award from ASJ in 2020, and the Itakura Prize Innovative Young Researcher Award from ASJ in 2021.

Hiroshi Sato
Researcher, Voice and Dialog Recognition Technology Group, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.
He received a B.E. and M.E. from the University of Tokyo in 2016 and 2018. He joined NTT in 2018 and has since been engaged in research, development, and practical application of speech-processing technologies. His research interests include speech enhancement, robust speech recognition, and speech dialog systems. He is a member of ASJ.

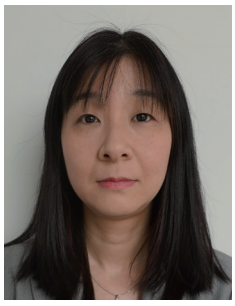
Yasunori Ohishi
Senior Research Scientist, Media Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received a Ph.D. from Nagoya University, Aichi, in 2009. Since joining NTT in 2009, he has been researching speech and audio signal processing. His research interests generally concern audio event detection, music information retrieval, and crossmodal learning with audio applications. He received the Awaya Prize Young Researcher Award from ASJ in 2014. He is a member of IEEE, ASJ, the Information Processing Society of Japan (IPSJ), and IEICE.

Keisuke Kinoshita
Distinguished Researcher, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received an M.E. and Ph.D. from Sophia University, Tokyo, in 2003 and 2010. After joining NTT Communication Science Labs in 2003, he has been engaged in fundamental research on various types of speech, audio, and music signal processing, including 1ch/multi-channel speech enhancement (blind dereverberation, source separation, noise reduction), speaker diarization, robust speech recognition, and distributed microphone array processing, and developed several innovative commercial software. He is an author or a co-author of more than 20 journal papers, 5 book chapters, more than 100 papers presented at peer-reviewed international conferences, and an inventor or a co-inventor of more than 20 Japanese patents and 5 international patents. He began serving as an associate editor of IEEE/ACM Transactions on Audio, Speech and Language Processing in 2021 and has been a member of IEEE SPS Audio and Acoustic Signal Processing Technical Committee (AASP-TC) since 2019. He served as the chief coordinator of the REVERB challenge (2014), editor of IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (from 2013 to 2017), and guest editor of EURASIP journal on advances in signal processing (2015). He was honored to receive the 2006 IEICE Paper Award, the 2010 ASJ Outstanding Technical Development Prize, the 2011 ASJ Awaya Prize, the 2012 Japan Audio Society Award, 2015 IEEE ASRU Best Paper Award Honorable Mention, and 2017 Maejima Hisoka Award. He is a member of IEEE, ASJ, and IEICE.

Tomohiro Nakatani

Senior Distinguished Researcher, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E., M.E., and Ph.D. from Kyoto University in 1989, 1991, and 2002. Since joining NTT as a researcher in 1991, he has been investigating speech-enhancement technologies for developing intelligent human-machine interfaces. He was a visiting scholar at Georgia Institute of Technology, USA, in 2005. He was a visiting assistant professor in the Department of Media Science, Nagoya University, from 2008 to 2018. He received the 2005 IEICE Best Paper Award, the 2009 ASJ Technical Development Award, the 2012 Japan Audio Society Award, the 2015 IEEE ASRU Best Paper Award Honorable Mention, and the 2017 Maejima Hisoka Award. He was a member of the IEEE SPS AASP-TC from 2009 to 2014 and has been a member of the IEEE SPS SL-TC since 2016. He served as an associate editor of the IEEE/ACM Transactions on Audio, Speech and Language Processing from 2008 to 2010, chair of the IEEE Kansai Section Technical Program Committee from 2011 to 2012, chair of the IEEE SPS Kansai Chapter from 2019 to 2020, a workshop co-chair of the 2014 REVERB Challenge Workshop, and general co-chair of the IEEE ASRU. He is a fellow of IEEE, and member of IEICE and ASJ.



Shoko Araki

Group Leader and Senior Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

She received a B.E. and M.E. from the University of Tokyo in 1998 and 2000, and Ph.D. from Hokkaido University in 2007. Since she joined NTT in 2000, she has been engaged in research on acoustic signal processing, array signal processing, blind source separation, meeting diarization, and auditory scene analysis. She was a member of the IEEE SPS AASP-TC from 2014 to 2019. She has been a board member of ASJ since 2017. She also served as a member of the organizing committee of the International Symposium on Independent Component Analysis and Blind Signal Separation (ICA) 2003, the International Workshop on Acoustic Signal Enhancement (IWAENC) 2003, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2007, the Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA) 2017, IEEE WASPAA 2017, IWAENC 2018, and the evaluation co-chair of the Signal Separation Evaluation Campaign (SiSEC) 2008, 2010, and 2011.

She received the 19th Awaya Prize from ASJ in 2001, the Best Paper Award of the IWAENC in 2003, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004 and 2014, the Academic Encouraging Prize from IEICE in 2006, the Itakura Prize Innovative Young Researcher Award from ASJ in 2008, The Young Scientists' Award of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2014, IEEE SPS Best Paper Award in 2014, and IEEE ASRU 2015 Best Paper Award Honorable Mention in 2015. She is a member of IEEE, IEICE, and ASJ.

Technique for Modulating the Tactile Sensation of Objects Using an Illusion

Takumi Yokosaka, Scinob Kuroki, and Shin'ya Nishida

Abstract

The human ability to determine the tactile textures of objects seems to be very stable: When someone rubs, holds, hits, or touches a stone, for example, she/he never confuses its tactile texture with that of fur or a sponge. However, the phenomenon known as the *velvet hand illusion* indicates that human perception of tactile texture can be easily distorted. This article describes a psychological study that examined the nature of this tactile illusion and presents a technique with which the perception of texture can be modulated using this phenomenon.

Keywords: haptics, tactile illusion, material perception

1. Learning from illusions

Compared with visual displays (monitors) and auditory displays (loudspeakers), it may seem that tactile displays have not reached a practical level, but their ability to represent tactile sensation is steadily improving. For example, the controller of a videogame console can simulate the tactile texture (how something feels to the touch) of a variety of objects by controlling a built-in oscillator. However, a technique capable of eliciting the perception of a tactile sensation, i.e., giving someone the impression that she/he is physically touching the object in question rather than receiving a tactile sensation via a controller, is yet to be established. When someone touches an object directly, his/her perception of the texture of that object is based not only on the kinetic sensation on the skin but also on various other physical properties, such as shape, irregularities on the surface, elasticity, heat conductivity, and moisture content. This is why it is difficult to fabricate a device that can flexibly represent the complex properties and states of an object. A clue to solving this problem is using a tactile illusion.

Visual illusions are well known, such as an object that ought to appear stationary looks as if it is moving. However, there are also tactile illusions, which are perceived via touch on the skin of the hands, for example. Tactile illusions can take a variety of forms depending on the properties that someone feels are different from the actual properties, such as shape, weight, movements, and tactile texture [1, 2]. Tactile illusions are important because they provide a clue as to how we estimate the properties of an object when we touch it. This article focuses on the *velvet hand illusion*, which is a tactile texture illusion. This illusion occurs when wires, such as tennis racket strings, are held between two hands and either the hands or the wires are moved back and forth, resulting in a tactile sensation that is strange and different from that of the wires (**Fig. 1**) [3]. The reason it is called the "velvet" illusion is due to the fact that the strange tactile sensation produced is similar to the sensation of touching velvet. This sensation provides a significant hint about tactile illusions. That is, in this illusion, a sensation like that of touching velvet, while completely at odds with the real object being touched, is produced directly on the skin of the palms. This illusion can be induced simply by moving the wires despite the fact none of the complex characteristics of the wires, such as shape, irregularities on the surface, and elasticity, are controlled. We surmised that a close examination of this phenomenon could provide a clue as to how to create a variety of tactile but illusory sensations directly on the skin of the hands.

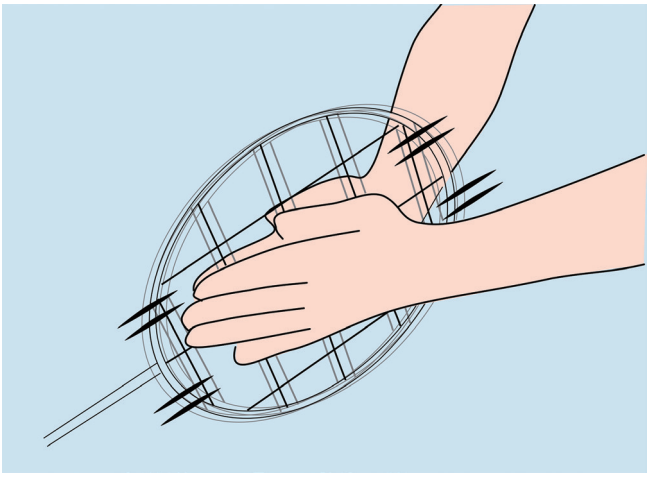


Fig. 1. The velvet hand illusion occurs when wires are sandwiched between both hands and either the hands or wires are moved back and forth in a rubbing motion.

2. What type of illusion is the velvet hand illusion?

When someone experiences the velvet hand illusion, she/he finds it difficult to describe precisely the strange tactile sensation she/he feels when the wires are moved. This makes it difficult to define the question that the study of the velvet hand illusion needs to focus on. That is, it is difficult to understand qualitatively and quantitatively what type of illusion the velvet hand illusion is and how it is caused. Therefore, we conducted an experiment to gain a clearer understanding of what kind of phenomenon the velvet hand illusion is [4]. In this experiment, the participants compared the tactile sensation when wires were moved between the two palms, namely, what they felt while the velvet hand illusion was occurring (**Fig. 2(a)**) and the tactile sensation produced with various commonly used materials (**Fig. 2(b)**). To identify how the tactile sensation varies when the intensity of the velvet hand illusion changes, we controlled the intensity of this illusion by presenting a variety of wire-related conditions, i.e., manipulating the distance between the wires, moving them vertically or horizontally, and touching them only with one hand (**Fig. 2(c)**). We also asked the participants to indicate the intensity of the velvet hand illusion for each condition.

The degree of similarity between the tactile sensation in the velvet hand illusion and the real texture of various materials was visualized on a two-dimensional (2D) space, as shown in **Fig. 3**. In other experiments in which the textures of various materials and those in the velvet hand illusion were evaluated in terms of roughness and hardness, we found that the horizontal axis of the 2D space corresponds to softness and warmth and the vertical axis to smoothness. The intensities of the illusory sensations evaluated under various wire-related conditions were not used in this analysis. Nonetheless, the different tactile sensations were plotted linearly from the condition of weak illusion intensity to that of strong illusion intensity. The correlation coefficient between the illusion intensity and horizontal axis was 0.75 and that between the illusion intensity and vertical axis was –0.93. It was found that the illusion was the weakest when the wires were touched with one hand (bottom right in Fig. 3). In this case, what was felt when touched most closely resembled the actual texture of a wire mesh. This seems a reasonable result considering that both the wires and a wire mesh are hard and linear materials. It was also found that the stronger the illusion, the more the tactile sensation shifts towards the upper left in Fig. 3, indicating that the tactile sensation gets softer and smoother. The illusion was the strongest when the wires were 75 mm apart, sandwiched between both hands, and moved back and forth. The tactile sensation was similar to the real texture of cloth or leather. These experimental results revealed that the velvet hand illusion is a phenomenon in which humans perceive that the texture of wires, which are inherently hard and rough, changes into something soft and smooth such as cloth or leather when the wires are moved between the

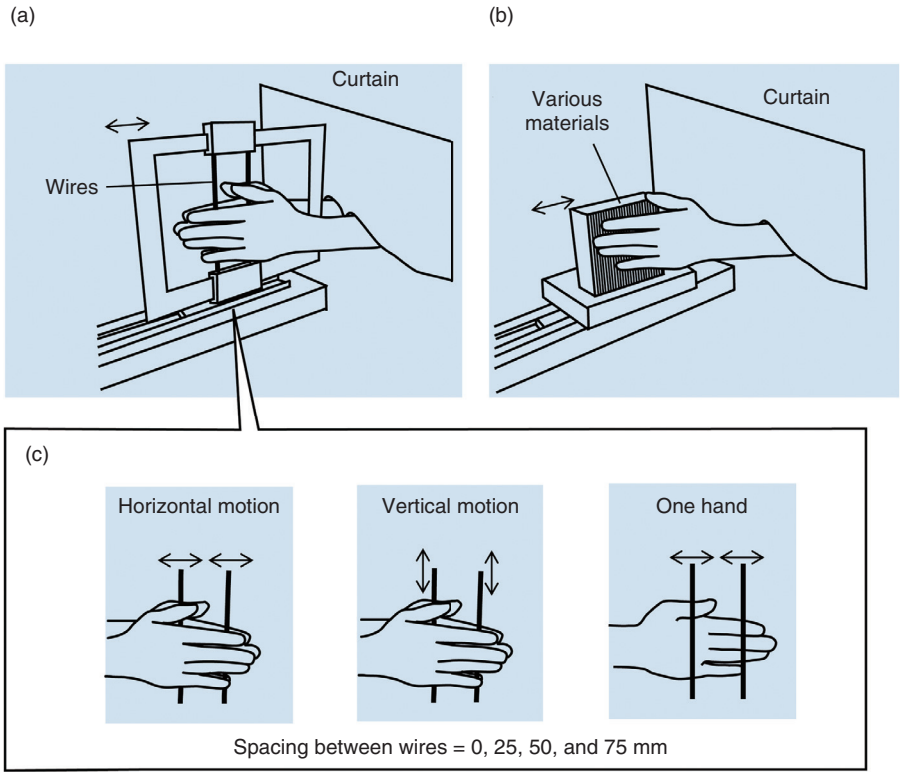


Fig. 2. The experimental setup to evaluate the tactile sensation of the velvet hand illusion [4]. (a) Presentation of wires. (b) Presentation of various materials. (c) A variety of wire-related conditions.

hands.

3. Extension of the illusion for any material

It has become clear that the velvet hand illusion is a phenomenon in which the perceived texture of directly touched wires changes. Can we apply this illusion to other materials besides wires? This is an important question to ask when studying potential techniques of presenting various tactile sensations. One obvious technique is to sandwich the wires between one hand and an object other than the other hand. However, there are several problems with this technique. For example, when wires are moved back and forth between a hand and object, the close contact between the hand and object can be lost, making the tactile sensation produced by the wires dominant. In addition, if the object has a rough surface, the wires can snag, making the movement bumpy. To solve these problems, we have taken into account the fact that the velvet hand illusion can also be produced by moving a thin board with a hole in it, instead of wires, between two finger pads [5]. This technique also allows the board to be sandwiched between two hands (**Fig. 4(a)**). It has also been found that the velvet hand illusion can be produced not only by moving the hole back and forth but also by rotating it (**Fig. 4(b)**). These findings suggest that it is not necessary to move the hole across the entire palm to produce the illusion. This led us to hypothesize that rubbing the hole with the outer part of the palm could produce the illusion in the central part of the palm confined by the edge. We have thus discovered a technique of manipulating the tactile sensation of a touched object by rotating a sheet of heavy paper held between a hand and object (**Fig. 4(c)**). We refer to this technique as the frame-rotation technique. This technique does not require anything to move a long distance between a palm and object. Also, because few edges exist in the rotating direction, they do not get caught on the object; thus, the paper can be moved smoothly.

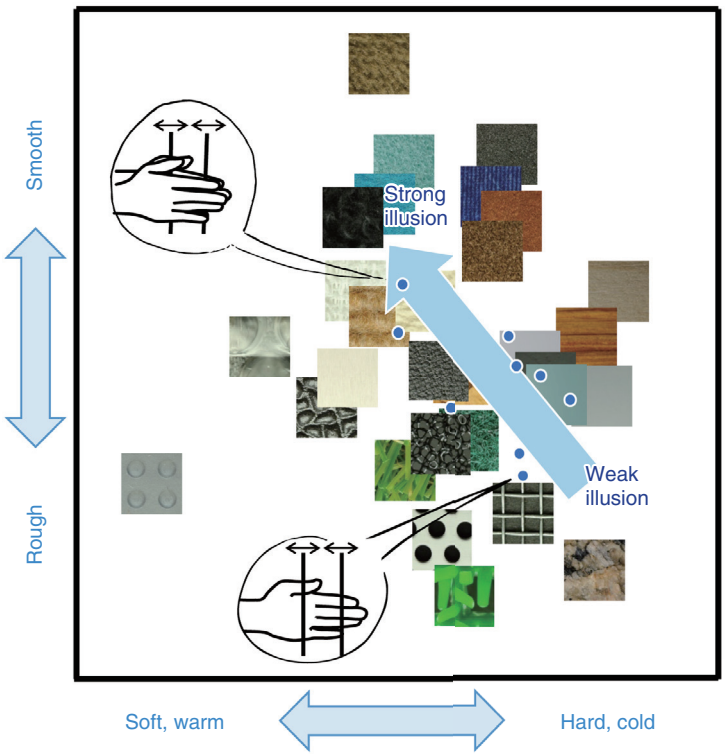


Fig. 3. Results of the tactile-texture evaluation experiment for the velvet hand illusion [4]. The blue dots indicate the conditions under which the wires are touched (whether the wires were touched with both hands or with one hand, whether the wires were moved horizontally or vertically, etc.).

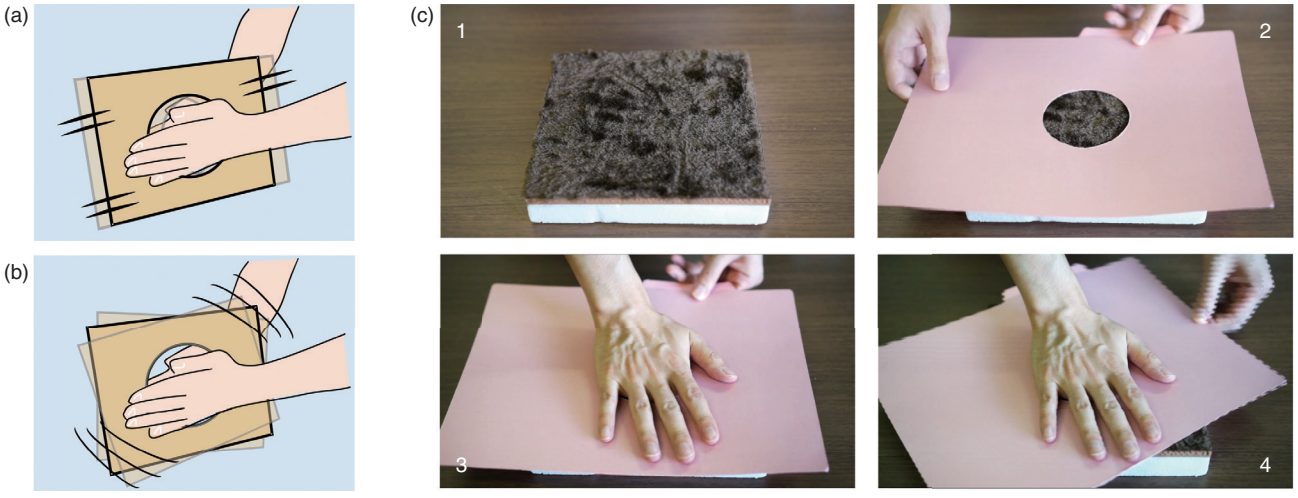


Fig. 4. Extension of the velvet hand illusion ((a) and (b)) and the frame-rotation technique (c).

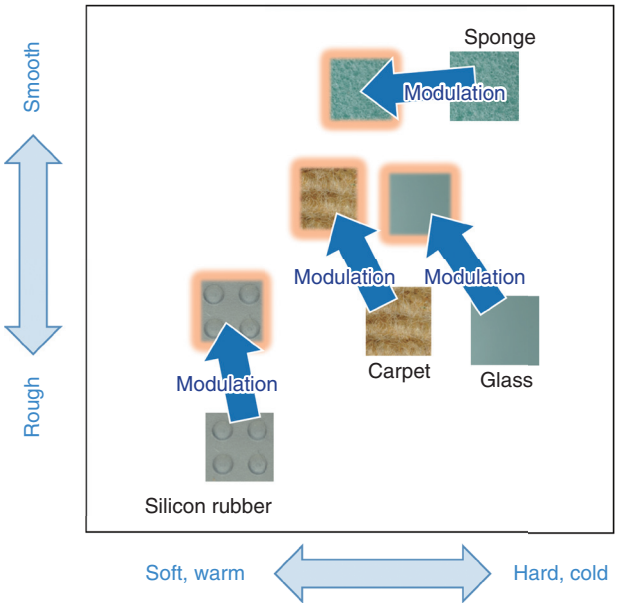


Fig. 5. Experimental results on the modulation of tactile sensation using the frame-rotation technique [6]. The dark blue arrows on the 2D space indicate how this technique changed the tactile impression of the respective object.

4. The perception of the tactile texture of an object can be made to be soft and smooth using the frame-rotation technique

We conducted experiments to ascertain whether this technique can make a range of objects feel soft and smooth [6]. Participants were asked to evaluate the textures of various materials and how similar the textures of the same materials were when the frame-rotation technique was applied. As in Fig. 3, the experimental results were visualized in a 2D space, as shown in **Fig. 5**. In other experiments in which the textures of these materials were evaluated in terms of roughness and hardness, we found that the horizontal axis in the 2D space is related to softness and warmth and the vertical axis to smoothness. For example, it was found that when the frame-rotation technique was applied to a rough and hard carpet, the perception of the carpet's texture changed to being softer and smoother. Thus, this technique makes the texture of various objects seem softer than they actually are.

As discussed above, we developed and evaluated a technique using the phenomenon in which the texture of an object can be made to seem softer and smoother, similar to the velvet hand illusion. The study of the brain mechanism that causes this illusion may enable us to discover hitherto unknown tactile-sensation processing mechanisms. It would also allow us to develop a technique for making the perception of an object's texture seem harder and rougher and one that allows the perception of softness and smoothness to be changed independently. A characteristic of this newly developed technique is that it requires no special device. Anyone can easily implement it using items found at home. Thus, anyone without specialist knowledge on the operation of a vibration-presentation or power-generation device will be able to easily control the perception of texture. For example, this technique would be easy to use when a designer wants to convey the tactile texture of a package to his/her clients or a sales clerk wants to demonstrate the effect of a product such as a softening agent.

References

[1] J. Watanabe, "Communication Research Focused on Tactile Quality and Reality," NTT Technical Review, Vol. 9, No. 11, 2011. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201111fa6.html

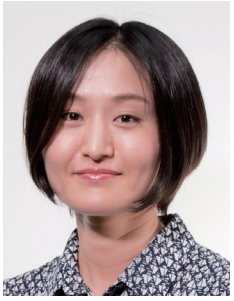
[2] S. Kuroki, "Towards Understanding Human Skin Sensations," NTT Technical Review, Vol. 18, No. 11, pp. 16–20, 2020. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202011fa2.html

[3] H. Mochiyama, A. Sano, N. Takesue, R. Kikuuwe, K. Fujita, S. Fukuda, K. Marui, and H. Fujimoto, "Haptic Illusions Induced by Moving Line Stimuli," Proc. of World Haptic Conference, pp. 645–648, Pisa, Italy, Mar. 2005.

[4] T. Yokosaka, S. Kuroki, and S. Nishida, "Describing the Sensation of the 'Velvet Hand Illusion' in Terms of Common Materials," IEEE Trans. Haptics, doi: 10.1109/TOH.2020.3046376.

[5] M. Ohno, T. Miyaoka, and M. Nakatani, “Two Hands Feel Smoother Than One: A Study on the Smoothness Magnitude Produced by the Velvet Slit Tactile Illusion,” EuroHaptics 2018, Pisa, Italy, June 2018.
[6] T. Yokosaka, Y. Suzuishi, and S. Kuroki, “Feel Illusory Texture through a Hole: Rotating Stimulus Modulates Tactile Sensation for Touched Object’s Surface,” EuroHaptics 2020, Leiden, Netherlands, Sept. 2020.

Takumi Yokosaka
Research Scientist, Sensory Representation Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories.
He received a B.E. and M.S. in engineering and information science from Osaka University in 2011 and 2013, and Ph.D. in information processing from Tokyo Institute of Technology in 2018. His research interests include perception and cognition shaped by body movement.



Scinob Kuroki
Senior Research Scientist, Sensory Representation Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories.
She received a Ph.D. in information science and technology from the University of Tokyo in 2011. Her research is focused on human tactile processing, particularly frequency perception, time perception, and motion perception.

Shin’ya Nishida
Visiting Senior Distinguished Scientist, Sensory Representation Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories and Professor, Graduate School of Informatics, Kyoto University.
He received a B.A., M.A., and Ph.D. in psychology from Kyoto University in 1985, 1987, and 1996 and joined NTT in 1992. His research focuses on visual motion perception, material perception, time perception, haptics, and multisensory integration. He is also interested in leveraging vision science for innovation of media technologies. He was/is on the editorial boards of Journal of Vision (from 2007), Vision Research (from 2008 to 2017), and Multisensory Research (from 2017). He was president of Vision Society of Japan (from 2014 to 2018) and is a member of Science Council of Japan (from 2017). He served as Rank Prize Lecturer at European Conference on Visual Perception 2017.

Routing and Spectrum Assignment Using Deep Reinforcement Learning in Optical Networks

Masayuki Shimoda and Takafumi Tanaka

Abstract

In future optical networks, effective use of spectral resources will be an issue as complexity increases due to diversified and dynamic requirements. In this article, we first give an overview of the routing and spectrum assignment (RSA) problem in elastic optical networks. We then introduce our novel RSA algorithm called Mask RSA, which allows for efficient route and spectral resource selection by deep reinforcement learning.

Keywords: routing and spectrum assignment, reinforcement learning, optical networks

1. Introduction

Internet traffic has been grown rapidly due to increasingly diversified traffic demands as a result of both video-streaming and cloud-computing services. To handle the changing requirements of future traffic in backbone networks, the concept of the elastic optical network (EON) was proposed [1]. EONs have a finely granular frequency grid (typically a multiple of either 6.25 or 12.5 GHz) to allocate the minimum frequency bandwidth to each channel, compared with the traditional flexed-grid (e.g., 50 GHz). EONs allocate a different number of frequency slots (FSs) to a connection request in accordance with the bandwidth demand for more efficient spectral utilization. However, non-uniform FS allocation results in spectral fragmentation that degrades spectral utilization. Therefore, efficient use of the spectral resources in EONs requires a routing and spectrum assignment (RSA) algorithm that can prevent spectral fragmentation.

Research into optical networking has yielded proposals based on RSA using deep reinforcement learning (DRL) [2] in EONs. The pioneering studies are on DeepRMSA [3] and its improved version [4], which determine the routing path among K shortest paths (KSPs), outperforming traditional RSA based on shortest path and KSP algorithms. These pioneering studies indicated that DRL-based RSA is promising. In this article, we introduce the basic concept of RSA in EONs followed by DRL-based RSA algorithms for efficient network planning in EONs. We then explain our proposed DRL-based RSA algorithm called Mask RSA along with its performance evaluation.

2. EONs

EONs have emerged as one of the most promising network technologies for next-generation optical networks. In simulation, an EON consists of nodes and links, and each link has FSs. FSs are represented as indexed list of FS status, e.g., available and occupied FSs are represented as 1 and 0, respectively, then FSs indicate a list [0,0,0,1,1,1,0,1,1]. Each connection request has a duration, and when the duration elapses, the FSs that were used are released.

Figure 1 illustrates an example of a simulated EON that consists of four nodes, A, B, C, and D, and four links, A-B, B-C, C-D, and D-A. Each link has eight FSs. Let us take the case of an incoming connection request from node A to node C with two FSs. To check which FSs are available to assign in the route A-B-C, FSs of links in the route are calculated by the bitwise AND operation; the FSs of A-B are

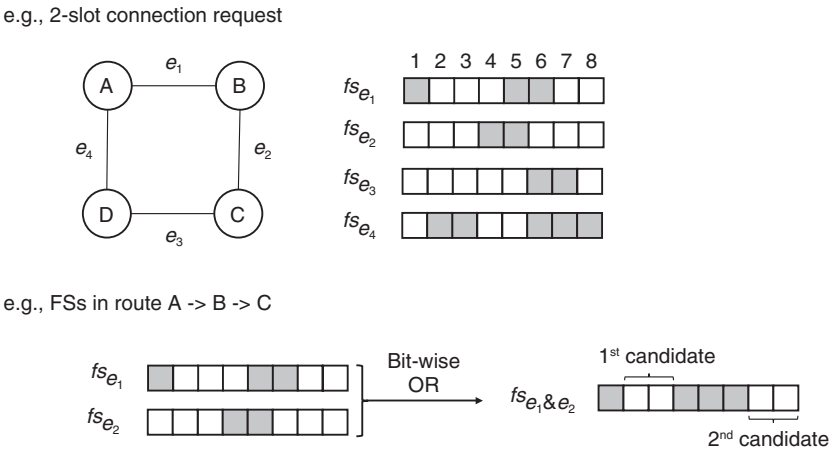


Fig. 1. Spectrum assignment example in an EON.

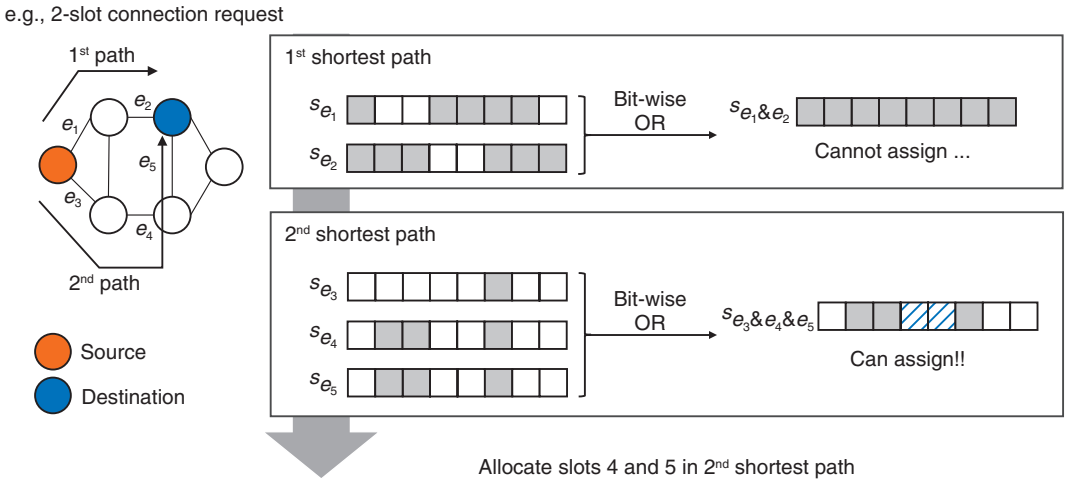


Fig. 2. Overview of KSP-FF algorithm.

[0,1,1,1,0,0,1,1] and B-C are [1,1,1,0,0,1,1,1], and the resulting FS status is [0,1,1,0,0,0,1,1]. In the route, assignable candidates are the 2nd and 3rd and 7th and 8th FSs. Like this example, a spectrum continuity constraint (i.e., same FSs should be used across links) exists, which ensures that all links in the end-to-end route use the same FSs. Spectrum assignment (SA) algorithms determine which candidates are used, as explained in the following section.

3. RSA

To give an overview of RSA, let us take an example of a well-known heuristic algorithm, the KSP and first fit (KSP-FF) algorithm. The KSP-FF algorithm solves the routing subproblem (by KSP) and SA subproblem (by FF) separately. It first determines a route then the FSs to be used. **Figure 2** shows an overview of determining which route and FSs.

In routing, the KSP-FF algorithm first precomputes a routing table that is an ordered list of the *K* routes between all pair of nodes. Next, when a connection request arrives, KSPs are obtained from the routing table by using the pair of source and destination nodes. Finally, a path where assignable FSs exist is searched for in order of shorter paths. When available FSs do not exist in *K* routes, the connection request is rejected, which is called blocking. After routing,

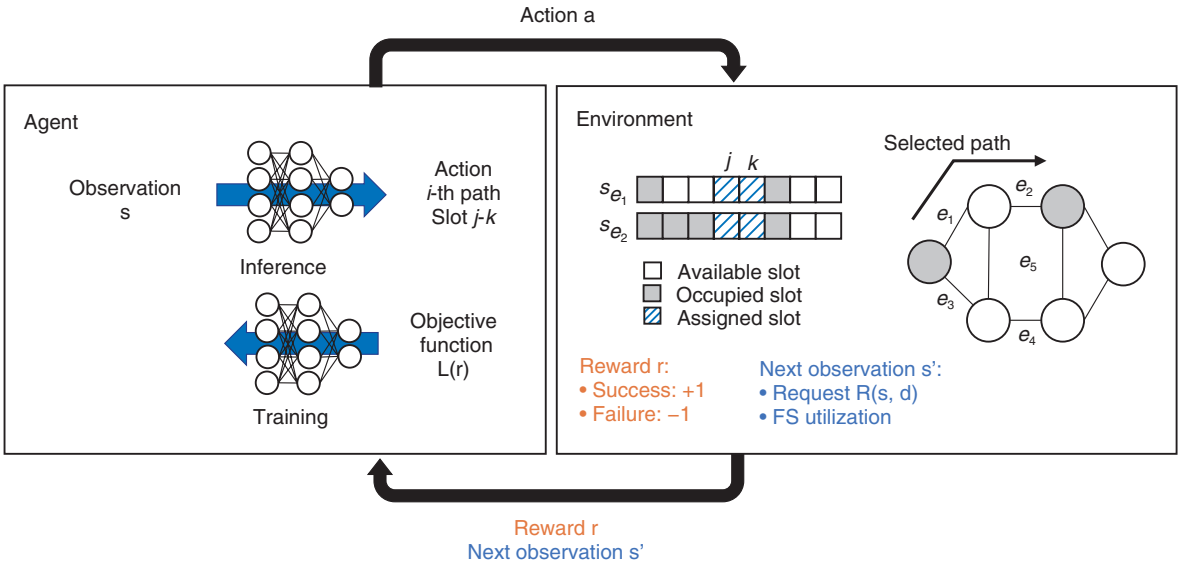


Fig. 3. Overview of DRL-based RSA.

spectrum assignment by FF is executed; in all assignable FSs, the lowest indexed FSs are selected. This assignment procedure packs existing connections into the smallest number of FSs, leaving a larger number of available FSs.

An efficient RSA algorithm can prevent spectral fragmentation. Let us take two cases in which FSs are [1,0,0,1,0,0,1] and [0,0,0,0,1,1,1] with a 2-FS request. In the first case, there are many spectral losses, making it impossible to assign the 2-FS request. The second case allocates FSs consecutively, so there is no spectrum loss and 2-FS requests can be allocated. Therefore, preventing spectral fragmentation can allocate more requests.

4. DRL-based RSA

An overview of dynamic RSA in the format of a well-known DRL modeling is shown in **Fig. 3**. At each time step, the agent takes an action, and the action space is pre-defined at the formulation phase. For example, when DRL is applied to the routing subproblem that selects one of the KSPs, the action space is {1, 2, ..., K}, the action of which is mapped to the corresponding shortest path. Next, the agent receives an observation and reward from the environment. Observation is the status of the current environment which includes a request and a status of FS utilization. Reward is a value that represents how good the action is. An agent takes actions on the basis of the observation, and parameters of the agent's action-decision function, i.e., a deep neural network (DNN) is updated to maximize the total number of rewards.

One of the key problems with DRL-based RSA is how to define the action space; assignable FSs vary time to time and depends on routing paths. **Figure 4** shows an example of the FS selection from action spaces proposed in a previous study [4]. The action space is {1, 2}, the action of which is mapped to assignable candidates. In case 1, the first candidate is 1–2 FSs, and the second one is 6–7 FSs. The DNN determines which candidate should be used to accommodate as many future connection requests as possible. Unlike case 1, the problem in case 2 is that the number of assignable candidates is less than the size of the action space. In this case, for actions that are not mapped to any assignable candidates, blocking is mapped; the blocking action is selected, and its connection request is rejected. Since this definition of action cannot be used to evaluate all assignable candidates properly when the size of the action space is less than the number of assignable candidates, a trained agent would be suboptimal. Therefore, efficient DRL-based RSA algorithms need to be flexible in accordance with the changing number of assignable candidates while avoiding spectral fragmentation.

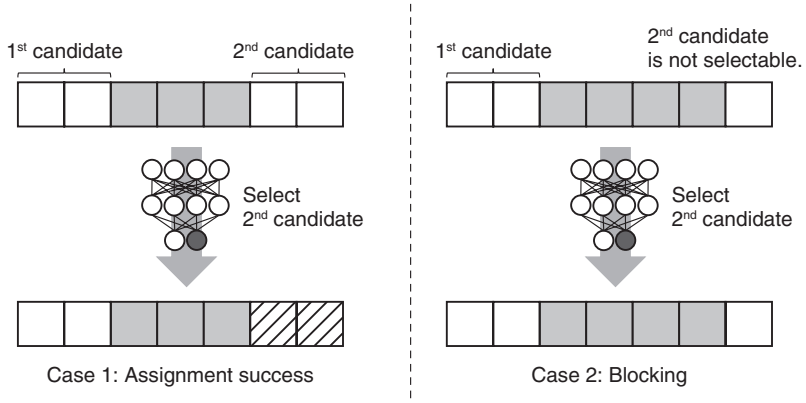


Fig. 4. Example of FS selection.

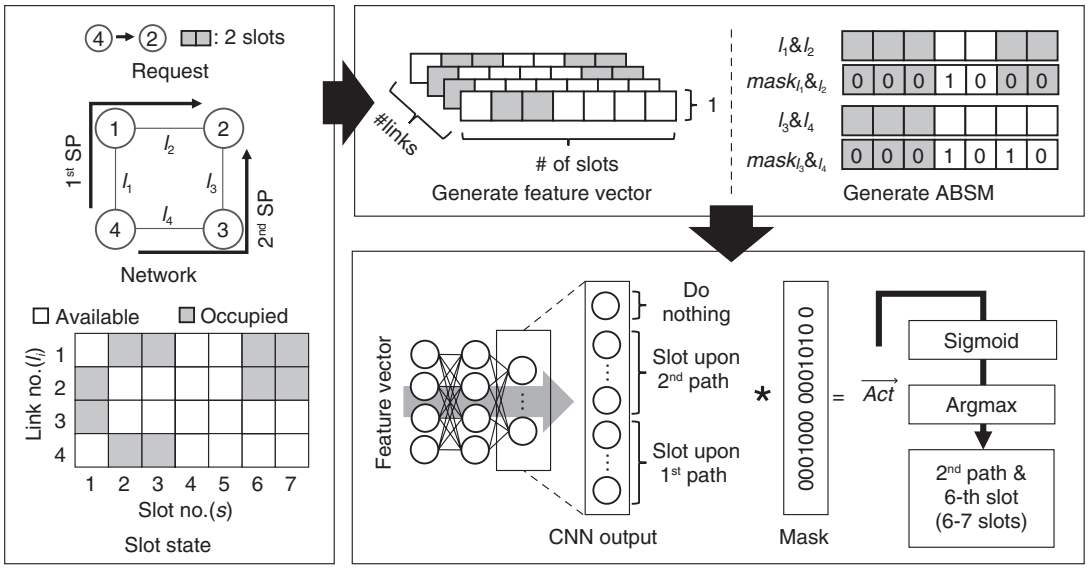


Fig. 5. Inference of Mask RSA.

5. Mask RSA

To handle the dynamic changes in the number of assignable candidates without spectral fragmentation, our Mask RSA masks unassignable candidates to take into account assignable ones. Mask RSA is based on our past study [5].

First, we explain the definition of an action space in Mask RSA. Mask RSA implements routing by selecting one of the KSPs, and SA is executed by selecting the first index of used FSs. Thus, an action space in Mask RSA is defined as $\{1, 2, ..., S \times K, S \times K + 1\}$, where S and K are the numbers of FSs and paths, respectively. The option of an action is do-nothing; if assignable resources do not exist in a KSP, take the action of do-nothing, which leads to blocking. For example, when the selected action number is 120, (1) do-nothing if $120 > S \times K$; otherwise (2) the selected path is $\lfloor 120/S \rfloor$ and the start index of used FSs is 120 *mod* S. This formulation makes an agent select a routing path and FSs concurrently.

Figure 5 gives an overview of Mask RSA inference. To handle dynamic changes in the number of assignable candidates, the masking approach is used. First, for each routing path, an assignable boundary slot mask ($\overrightarrow{ABSM}$) is generated as a vector with its

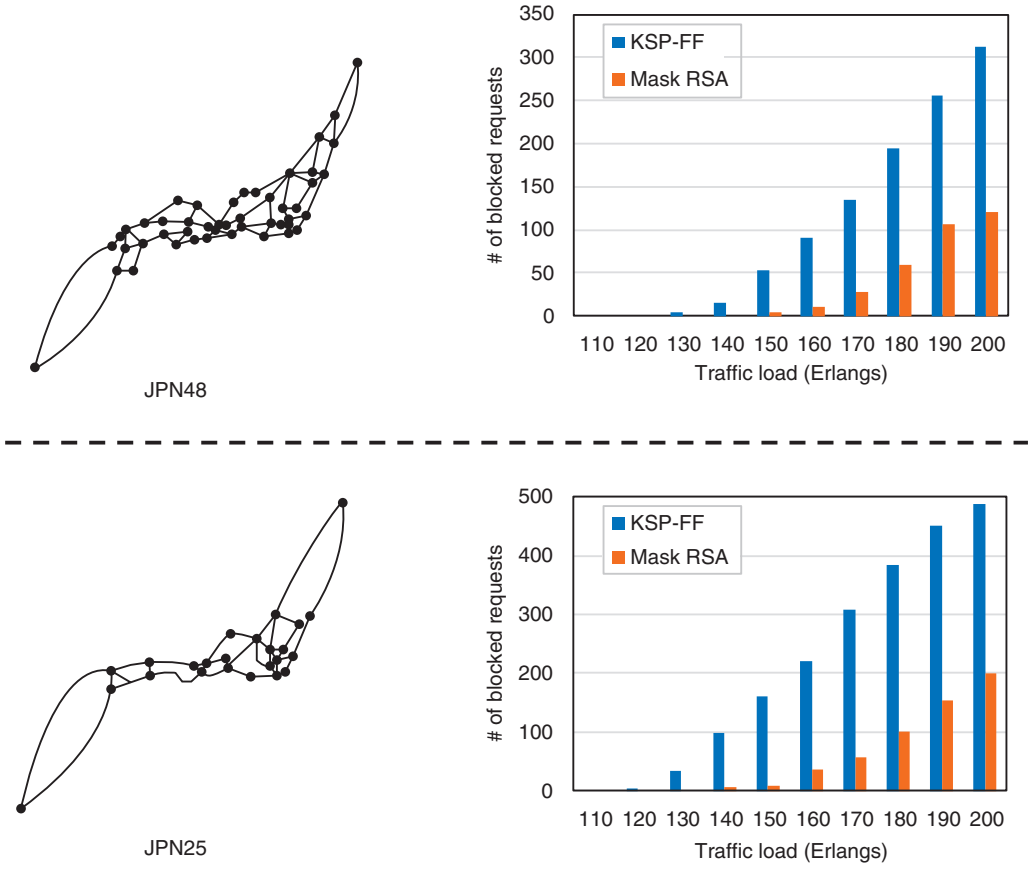


Fig. 6. Simulation results.

assignable position at a boundary of 1; otherwise, 0. An ABSM can prevent spectral fragmentation since it takes into account only boundaries. Next, a do-nothing mask is created; 0 when assignable candidates exist among KSPs, otherwise 1. Finally, both K ABSMs and do-nothing mask are concatenated to generate an RSA assignable mask ($\overrightarrow{RSA2M}$). Let *softmax*() and *argmax*() be softmax and argmax functions, respectively. The mapping function from convolutional neural network (CNN) output to an action is written as

$$Action\ no. = argmax\left(softmax(\overrightarrow{out} \circ \overrightarrow{RSA2M})\right),$$

where $\overrightarrow{out}$ is a DNN output vector, and $\circ$ is the Hadamard product. This mask prevents both the examination of unassignable choices and spectral fragmentation.

6. Demonstration

We evaluated the performance of Mask RSA by comparing it with the KSP-FF algorithm through simulations. The simulations involved a dynamic traffic scenario in which requests were generated on the basis of a Poisson process following a uniform traffic distribution. The average arrival rate and service duration for training were 10 and 12, respectively. The requested FS width was randomly selected in the range of 1 to 8. The tested networks were JPN25 (25 nodes and 43 links) and JPN48 (48 nodes and 82 links) [6]. Each FS was set to be 12.5 GHz, and each link had 320 FSs.

Figure 6 shows request blocking probabilities versus traffic loads from 110 to 200 Erlangs. In both networks, our Mask RSA outperformed the KSP-FF algorithm even when the traffic load differed from that used in training. These simulations in two topologies showed that the masking approach enabled efficient RSA regardless of the network size or traffic load used for training; accordingly, Mask RSA outperformed the KSP-FF algorithm.

7. Further studies

In this article, we introduced the application of DRL to dynamic optical network planning in EONs. The proposed DRL-based RSA algorithm, Mask RSA, is enhanced with domain-specific knowledge, outperforming heuristic algorithms. We plan to expand our study to more complicated networks, e.g., multi-core, multi-layer networks, and impairment-aware RSA, to enable more efficient network planning.

References

[1] O. Gerstel, M. Jinno, A. Lord, and S. J. B. Yoo, "Elastic Optical Networking: A New Dawn for the Optical Layer?", IEEE Commun. Mag., Vol. 50, No. 2, pp. s12–s20, Feb. 2012. doi: 10.1109/MCOM.2012.6146481

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level Control through Deep Reinforcement Learning," Nature, 518, pp. 529–533, Feb. 2015. https://doi.org/10.1038/nature14236

[3] X. Chen, J. Guo, Z. Zhu, R. Proietti, A. Castro, and S. J. B. Yoo, "Deep-RMSA: A Deep-reinforcement-learning Routing, Modulation and Spectrum Assignment Agent for Elastic Optical Networks," Proc. of 2018 Optical Fiber Communications Conference and Exposition (OFC), San Diego, CA, USA, Mar. 2018.

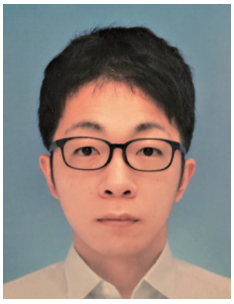
[4] X. Chen, B. Li, R. Proietti, H. Lu, Z. Zhu, and S. J. B. Yoo, "DeepRMSA: A Deep Reinforcement Learning Framework for Routing, Modulation and Spectrum Assignment in Elastic Optical Networks," J. Lightw. Technol., Vol. 37, No. 16, pp. 4155–4163, Aug. 2019.

[5] M. Shimoda and T. Tanaka, "Deep Reinforcement Learning-based Spectrum Assignment with Multi-metric Reward Function and Assignable Boundary Slot Mask," 2021 Opto-Electronics and Communications Conference (OECC), July 2021.

[6] S. Arakawa, T. Sakano, Y. Tsukishima, H. Hasegawa, T. Tsuritani, Y. Hirota, and H. Tode, "Topological Characteristic of Japan Photonic Network Model," IEICE Tech. Rep., Vol. 113, No. 91, PN2013-2, June 2013.

Masayuki Shimoda
Employee, NTT Network Innovation Laboratories.

He received a B.E. and M.E. in engineering from Tokyo Institute of Technology in 2018 and 2020. In 2020, he joined NTT Network Innovation Laboratories. He is a member of the Institute of Electrical and Electronics Engineers (IEEE).



Takafumi Tanaka
Research Engineer, NTT Network Innovation Laboratories.

He received a B.E. in electrical engineering and M.S. in informatics from the University of Tokyo in 2007 and 2009. Since joining NTT Network Innovation Laboratories in 2009, he has been engaged in R&D of next-generation photonic network architecture, planning, and control technologies. He received the IEEE JC Young Engineer Award from the IEEE Japan Chapter in 2012.

Next-generation Metallic Access Technologies and Their Standardization Activities

Yoshihiro Kondo and Noriyuki Araki

Abstract

This article introduces certain next-generation metallic access technologies and the standardization activities relating to them under study in the International Telecommunication Union's Telecommunication Standardization Sector (ITU-T) Study Group 15 with a strong focus on G.fast ("G" stands for ITU-T G series of recommendations and "fast" stands for fast access to subscriber terminals) and MGfast (multi-gigabit fast access to subscriber terminals) that can provide optical-fiber-grade ultrahigh-speed transmission services using pre-installed metallic cables in the existing infrastructure. MGfast, which is expected to be a next-generation metallic access technology, targets a transmission rate of 5 to 10 Gbit/s (uplink and downlink combined) over twisted-pair cables and coaxial cables.

Keywords: MGfast, G.fast, access

1. Development of metallic access networks

Question 4, which is responsible for the standardization of metallic access technologies, in the International Telecommunication Union's Telecommunication Standardization Sector (ITU-T) Study Group 15 (SG15) started to develop digital subscriber line (DSL)-related standards in 1998. Asymmetric DSL (ADSL) and very high-speed DSL (VDSL) technologies and standards have been developed by 2010 and provide capabilities supporting Internet access for home users. These technologies have supported the demand for high-speed Internet not only in Japan but also worldwide. Since then, network configurations that bring optical network termination closer to customer premises equipment (CPE), such as fiber to the cabinet (FTTC), fiber to the distribution point (FTTdp), fiber to the building (FTTB), and fiber to the home (FTTH), have been adopted due to the lower cost of optical fibers. One of these configurations connects optical fibers from the central office to the distribution point (DP) then to the CPE using G.fast ("G" stands for ITU-T G series of recommendations and "fast" stands for fast access to subscriber terminals). DPs are located in manholes, utility poles, and basements of apartment buildings, depending on the service provider. G.fast, included in **Fig. 1**, provides a transmission rate of 2 Gbit/s (uplink and downlink combined) over a 50-m twisted-pair cable. MGfast (multi-gigabit fast access to subscriber terminals), the next-generation metallic access technology described in detail in this article, targets a transmission rate of 5 Gbit/s (uplink and downlink combined) over a 30-m twisted-pair cable.

1.1 Ultrahigh-speed access technology G.fast

G.fast was originally standardized in 2014 (G.9700/G.9701) then revised in 2019 to achieve Gbit/s-level transmission rates with substantial expansion of frequency bands (adoption of 106- and 212-MHz profiles) and has the additional functionality of changing the transmission rates of uplink and downlink directions using the time division duplex (TDD) method. Unlike conventional ADSL and VDSL, which use the frequency division duplex (FDD) method, the TDD method allows easy control of the transmission rates in uplink and downlink directions, as shown in the frame structure in **Fig. 2**.

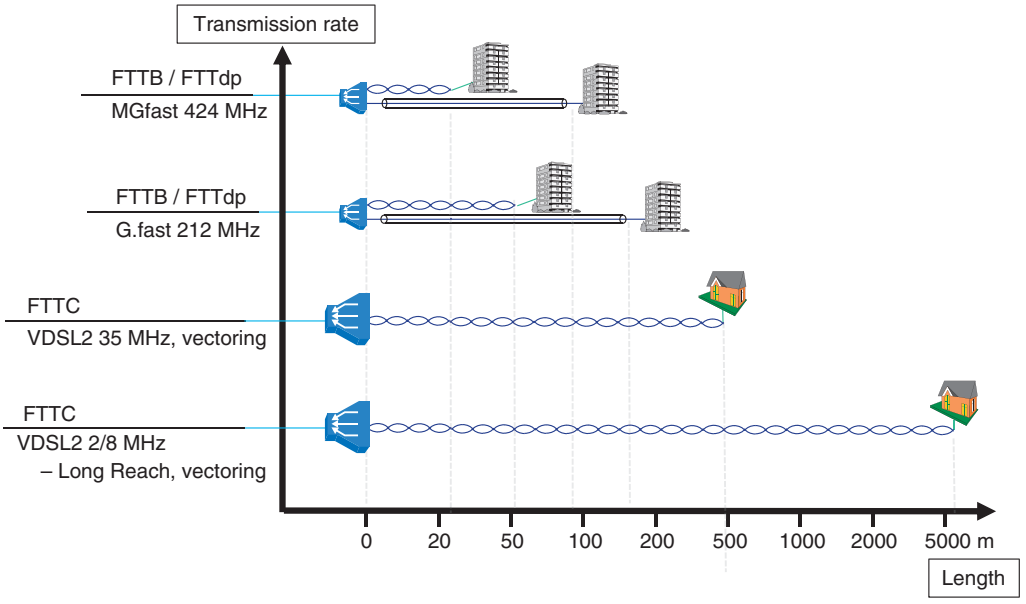


Fig. 1. Metallic access network solutions.

There are also the following features of specific profiling, which are also listed in **Table 1**:

- Transmission over coaxial cables in addition to twisted-pair cables
- Expansion of the transmission distance by increasing the transmission signal power
- Reverse power feeding (power-feeding capability from CPE to distribution point unit (DPU))

It should be mentioned that G.fast (106-MHz profile) is provided in Japan for commercial services for multi-dwelling units.

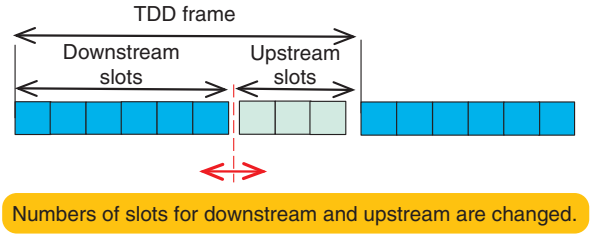


Fig. 2. Frame structure in G.fast.

1.2 Multi-gigabit ultrahigh-speed access technology MGfast

While G.fast was being deployed commercially, it was agreed at the ITU-T SG15 meeting in June 2017 to start developing standards for next-generation ultrafast access technology to achieve further enhancements in functionality and performance. Requirements proposed by service providers in Europe and the U.S. brought about the study of MGfast targeting transmission rates of 5 to 10 Gbit/s. As was the case with G.fast, reliable ultrahigh-speed transmission services of optical-fiber grade are required for next-generation access networks. In transmission systems using metallic cables, signals from neighboring lines result in interference, and how to eliminate such noise (far-end crosstalk noise and near-end crosstalk noise) is a major issue. In DSL technology, which uses a multi-carrier modulation scheme to achieve higher speeds, the subcarrier spacing and number of subcarriers are important system parameters as well as the frequency band used to complete subcarrier-by-subcarrier processing. Although it is effective to increase the number of subcarriers to achieve higher speed, the computational complexity increases accordingly; thus, the subcarrier spacing is generally determined to limit the number of subcarriers. As shown in **Table 2**, the subcarrier spacing is 51.375 kHz and the number of subcarriers is 2048 for G.fast (106-MHz profile). MGfast (424-MHz profile), which is described later, has the same subcarrier spacing, but the number of subcarriers is significantly increased to 8192. In MGfast (848-MHz profile), which is expected to be standardized, the subcarrier spacing is anticipated to

Table 1. Profile specifications in G.fast.

G.fast profiles	106a	106b	212a	106c	212c
Type of cables	Twisted-pair cables			Coaxial cables	
Maximum transmit power	+4 dBm	+8 dBm	+4 dBm	+2 dBm	+2 dBm
Precoding for vectoring	Linear coding			N/A	
Transmission rate with uplink and downlink combined	1 Gbit/s	1 Gbit/s	2 Gbit/s	1 Gbit/s	2 Gbit/s
Minimum frequency	2.2 MHz				
Maximum frequency	106 MHz	106 MHz	212 MHz	106 MHz	212 MHz
Limit PSD mask (LPM)	LPM_106	LPM_106 LPM_106high (downlink)	LPM_212	LPM_106	LPM_212

PSD: power spectral density

Table 2. System parameters for high speed copper access technologies.

DSL technology	Transmission scheme	Frequency band	Subcarrier spacing	Symbol rate	# of subcarriers	Transmission rate	ITU-T standard
VDSL2	FDD	17 MHz	4.3125 kHz	4 kHz	4096	100 Mbit/s	G.993.2
G.fast (106)	TDD	106 MHz	51.375 kHz	48 kHz	2048	1 Gbit/s	G.9700 G.9701
G.fast (212)	TDD	212 MHz	51.375 kHz	48 kHz	4096	2 Gbit/s	
MGfast (424)	FDX	424 MHz	51.375 kHz	48 kHz	8192	5 Gbit/s	G.9710/G.9711

be double.

A full duplex (FDX) method is also effective to increase the transmission rate. However, to avoid interference from near-end crosstalk noise and echo signals, VDSL uses the FDD method, while G.fast uses the TDD method. In this context, MGfast, a newly developed next-generation technology, is aimed at doubling the transmission rate by enabling full-duplexing in the single-pair while eliminating the effects of near-end crosstalk noise and echo signals.

2. Next-generation metallic access technology: MGfast standardization activities

This section describes the technical specifications and standardization status of MGfast, which has been studied in ITU-T SG15. MGfast is a technology to achieve even higher transmission rates while maintaining compatibility with G.fast for easy migration from G.fast systems. MGfast is mainly applicable to apartment buildings.

2.1 Objective of MGfast standardization

Standardization have been undertaken since 2017 with the following objectives:

- Provision of multi-gigabit access technology using twisted-pair cables for plain old telephone services (POTS) and coaxial cables for television services as transmission media, both of which are existing infrastructure
- Provision of symmetric and asymmetric communications with a combined transmission rate of up to 8 Gbit/s uplink and downlink using a frequency bandwidth of up to 424 MHz to achieve optical fiber grade high performance (extension of optical fiber)
- Removal of far-end and near-end crosstalk noise over multiple lines

2.2 Features of MGfast

The following features have been specified to enable easy migration from G.fast systems:

- Uplink and downlink transmission rates changeable by the operational mode of the TDD method
- Impulse noise removal by retransmission processing
- Frequency notch capability (use of the frequency bands in consideration of other services)
- Vectoring capability to eliminate far-end crosstalk noise

- Reverse power-feeding capability

The following new features are specified for the next-generation MGfast technology:

- The 424-MHz profile using frequency bandwidth up to 424 MHz is specified to achieve transmission rates of up to 5 Gbit/s in one direction. Although the specifications are optimized for operation over twisted-pair cables of 30 to 100 m, the specifications allow operation over twisted-pair cables and coaxial cables up to 400 m.
- The FDX method is specified to achieve almost twice the transmission rate over coaxial cables, while reducing the transmission delay at the same time. However, the simultaneous transmission of uplink and downlink signals requires removal of near-end crosstalk noise and echo signals.
- In addition to the forward error correction (FEC) combined trellis coded modulation and Reed-Solomon (RS), which is specified in G.fast, an advanced FEC combining low-density parity check/probabilistic constellation shaping and RS is newly specified. The adoption of this scheme also contributes to the improvement in the transmission rate.
- Quality-of-service classes based on the delay are specified separately for the uplink and downlink directions. This makes it possible to provide low-latency services.
- The point-to-multipoint configuration, in which multiple CPEs are connected to the same line, enables MGfast to meet the diversified needs of home users, unlike conventional VDSL and G.fast. It can change the set of subcarriers serving any particular CPE, which addresses a dynamic redistribution of bandwidth among multiple CPEs depending on traffic demands. Since multiple CPEs are connected to the same line, the Institute of Electrical and Electronics Engineers (IEEE) 802.1X authentication is included to ensure security.

The MGfast related standards consist of the following:

- ITU-T G.9710 published in February 2020: regulatory specifications, including frequency-related and power-spectrum-density-related issues
- ITU-T G.9711 approved in April 2021: system and physical layer specifications for MGfast
- ITU-T G.997.3 approved in April 2021: physical layer OAM (operations, administration and maintenance) specifications

3. Considerations for the future

System development based on the MGfast specifications approved at the ITU-T SG15 meeting held in April 2021 is expected to be accelerated. For service providers who do not have sufficient optical fiber assets, one of the options would be to provide high-performance services of optical-fiber grade using existing metallic cables. The standardization of the next-generation metallic access technologies described in this article is expected to continue to be studied to satisfy the demand for higher speed services by effectively using existing facilities. The following topics have been discussed in the standardization meetings as future themes for metallic access technology.

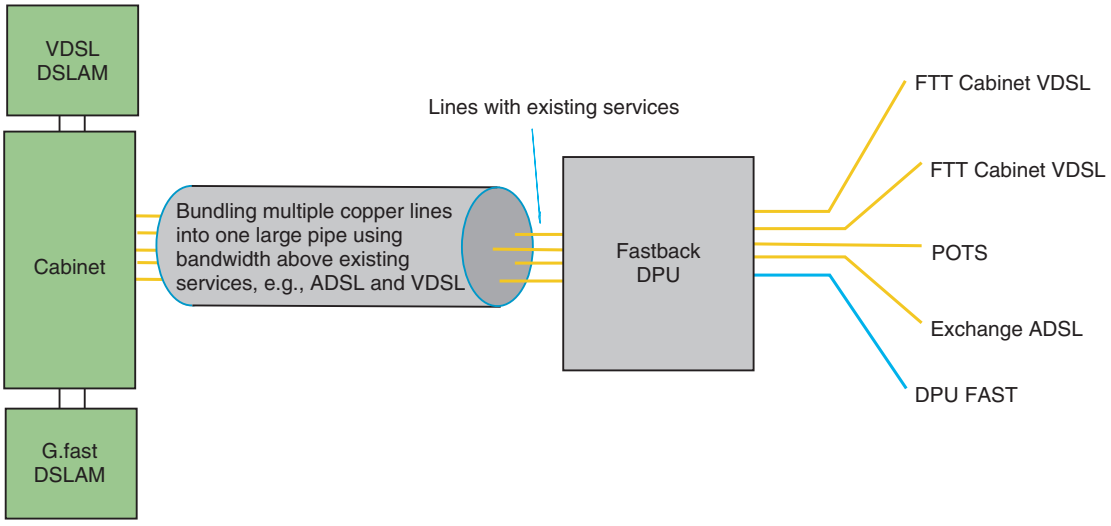
3.1 Specifications regarding MGfast (848-MHz profile)

The MGfast system for coaxial cables uses a frequency of up to 848 MHz. The doubling of the frequency band, combined with FDX technology, will make it possible to achieve transmission rates of up to 10 Gbit/s in one direction. It will be necessary to study the details of system parameters such as subcarrier spacing and the number of subcarriers, as explained earlier.

3.2 Bundling of multiple G.fast lines (G.fastback)

A project called G.fastback is under consideration as a technology to provide a transmission path between a cabinet and DPU as a single high-capacity line by combining multiple G.fast lines. It should be possible to use multiple metallic cables as an alternative to optical fiber cables with G.fastback technology. As application examples, this configuration would be connected in tandem in a number of stages to increase the distance, while in another example of G.fastback, multiple G.fast lines are bundled where G.fast signals are frequency multiplexed over the existing service lines such as ADSL and VDSL shown in **Fig. 3**. In both cases, the objective is to provide reliable and higher data-rate operations while effectively using metallic cables. Regarding G.fastback, there are several issues to be solved, which include how to eliminate near-end crosstalk noise in DPUs and signal synchronization between lines when multiple lines are connected.

In this article, the next-generation metallic access technologies and standardization activities being



DSLAM: digital subscriber line access multiplexer

Fig. 3. Deployment scenario of G.fastback.

developed in ITU-T SG15 were described, which can provide ultrahigh-speed transmission operations of optical-fiber grade by using existing metallic cables specified in G.9700 series standards, as an extension of optical fiber in FTTdp and FTTB configurations. It will be necessary to contribute further to the standardization activities mentioned in this article and to cooperate with service providers in Europe and the U.S. to use existing network facilities effectively and economically.

Yoshihiro Kondo
Senior Expert Engineer, IOWN Business Development Business Unit, Network Innovation Business Headquarters, NTT Advance Technology Corporation.
He received an M.S. and Ph.D. in electrical engineering from Northwestern University in Evanston, Illinoi, USA, in 1990 and 1994, with signal processing major. After working at OKI Electric Industry Co., Ltd. for developing optical access network systems, he joined NTT Advance Technology in 2006. Since then, he has been working on standardization activities in access networking systems, home networking systems, and some other areas. He is a member of IEEE.

Noriyuki Araki
Senior Manager, Standard Strategy, Research and Development Planning Department, NTT.
He received a B.E. and M.E. in electrical and electronic engineering from Sophia University, Tokyo, in 1993 and 1995. He joined NTT Access Network Service Systems Laboratories in 1995, where he researched and developed operation and maintenance systems for optical fiber cable networks. He has been contributing to standardization efforts in ITU-T SG6 since 2006. He was the rapporteur of Question 6 of ITU-T SG6 from 2006 to 2008 and the rapporteur of Question 17 of ITU-T SG15 from 2008 to 2012. He also served as the chairman of the ITU-T Focus Group on Disaster Relief Systems and Network Resilience and Recovery. He has been the vice-chairman of ITU-T SG15 since 2013. He also contributes to the activities of International Electrotechnical Commission (IEC) Technical Committee 86 (fiber optic systems). He received the ITU-AJ award from the ITU Association of Japan in 2017. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE).

External Awards

Specially Selected Paper
Winners: Takashi Koide, Daiki Chiba, Mitsuaki Akiyama, NTT Secure Platform Laboratories; Katsunari Yoshioka, Tsutomu Matsumoto, Yokohama National University
Date: May 19, 2021
Organization: Information Processing Society of Japan (IPSJ)

For "Understanding the Fake Removal Information Advertisement Sites."
Published as: T. Koide, D. Chiba, M. Akiyama, K. Yoshioka, and T. Matsumoto, "Understanding the Fake Removal Information Advertisement Sites," Journal of Information Processing, Vol. 29, pp. 392–405, May 2021.

PRMU Research Encouragement Award
Winners: Kazuki Adachi and Shin'ya Yamaguchi, NTT Software Innovation Center
Date: May 21, 2021
Organization: The Institute of Electronics, Information and Communication Engineers (IEICE) Technical Committee on Pattern Recognition and Media Understanding (PRMU)

For "Improving Accuracy on Biased Datasets via Explanations of Deep Neural Networks."
Published as: K. Adachi and S. Yamaguchi, "Improving Accuracy on Biased Datasets via Explanations of Deep Neural Networks," IEICE Tech. Rep., Vol. 120, No. 409, PRMU2020-93, pp. 139–144, Mar. 2021.

Best Paper Award
Winners: Yoshihiro Ogiso, Josuke Ozaki, Yuta Ueda, NTT Device Innovation Center; Hitoshi Wakita, NTT Device Technology Laboratories; Shigeru Kanazawa, Mitsuteru Ishikawa, NTT Device Innovation Center
Date: June 3, 2021
Organization: IEICE

For "Ultra-high Bandwidth and Low Drive Voltage InP-based IQ Optical Modulator for 100-GBd Class Optical Transmitter."
Published as: Y. Ogiso, J. Ozaki, Y. Ueda, H. Wakita, S. Kanazawa, and M. Ishikawa, "Ultra-high Bandwidth and Low Drive Voltage InP-based IQ Optical Modulator for 100-GBd Class Optical Transmitter," IEICE Trans. Electron., Vol. J103-C, No. 1, pp. 61–68, Jan. 2020 (in Japanese).

Encouraging Award
Winner: Rintaro Harada, NTT Access Network Service Systems Laboratories
Date: September 10, 2021
Organization: IEICE Technical Committee on Communication Systems

For "A Study on Optical Access Systems for 6G Radio Access Networks."
Published as: R. Harada, H. Ujikawa, N. Shibata, S. Kaneko, and J. Terada, "A Study on Optical Access Systems for 6G Radio Access Networks," IEICE Tech. Rep., Vol. 120, No. 107, CS2020-17, pp. 13–16, 2020.

Papers Published in Technical Journals and Conference Proceedings

Variational Secure Cloud Quantum Computing
Y. Shingu, Y. Takeuchi, S. Endo, S. Kawabata, S. Watabe, T. Nikuni, H. Hakoshima, and Y. Matsuzaki
arXiv:2106.15770, June 2021.

Variational quantum algorithms (VQAs) have been considered to be useful applications of noisy intermediate-scale quantum (NISQ) devices. Typically, in the VQAs, a parametrized ansatz circuit is used to generate a trial wave function, and the parameters are optimized to minimize a cost function. On the other hand, blind quantum computing (BQC) has been studied in order to provide the quantum algorithm with security by using cloud networks. A client with a limited ability to perform quantum operations hopes to have access to a quantum computer of a server, and BQC allows the client to use the server's computer without leakage of the client's information (such as input, running quantum algorithms, and output) to the server. However, BQC is designed for fault-tolerant quantum computing, and this requires many ancillary qubits, which may not be suitable for NISQ devices. Here, we propose an efficient way to implement the NISQ computing with guaranteed security for the client. In our architecture, only $N+1$ qubits are required, under an assumption that the form of ansatzes is known to the server, where N denotes the necessary number of the qubits in the original NISQ algorithms. The client only performs single-qubit measurements on an ancillary qubit sent from the server, and the measurement angles can specify the parameters for the ansatzes of the NISQ algorithms. No-signaling principle guarantees that neither parameters chosen by the client nor the

outputs of the algorithm are leaked to the server. This work paves the way for new applications of NISQ devices.

The Unicellular Red Alga *Cyanidioschyzon merolae*, an Excellent Model Organism for Elucidating Fundamental Molecular Mechanisms and Their Applications in Biofuel Production

I. Pancha, K. Takaya, K. Tanaka, and S. Imamura

Plants, Vol. 10, No. 6, 1218, June 2021.

Microalgae are considered one of the best resources for the production of biofuels and industrially important compounds. Various models have been developed to understand the fundamental mechanism underlying the accumulation of triacylglycerols (TAGs)/starch and to enhance its content in cells. Among various algae, the red alga *Cyanidioschyzon merolae* has been considered an excellent model system to understand the fundamental mechanisms behind the accumulation of TAG/starch in the microalga, as it has a smaller genome size and various biotechnological methods are available for it. Furthermore, *C. merolae* can grow and survive under high temperature (40°C) and low pH (2–3) conditions, where most other organisms would die, thus making it a choice alga for large-scale production. Investigations using this alga has revealed that the target of rapamycin (TOR) kinase is involved in the accumulation of carbon-reserved molecules, TAGs, and starch. Furthermore, detailed molecular mechanisms of the role of TOR in controlling the accumulation of TAGs and starch were uncovered via omics analyses. Based on these findings, genetic engineering of the key gene and proteins resulted in a drastic increment of the amount of TAGs and starch. In addition to these studies, other trials that attempted to achieve the TAG increment in *C. merolae* have been summarized in this article.

Articulatory Compensation for Low-pass Filtered Formant-altered Auditory Feedback

Y. Uezu, S. Hiroya, and T. Mochida

The Journal of the Acoustical Society of America, Vol. 150, No. 1, pp. 64–73, July 2021.

Auditory feedback while speaking plays an important role in stably controlling speech articulation. Its importance has been verified in formant-altered auditory feedback (AAF) experiments where speakers utter while listening to speech with perturbed first (F1) and second (F2) formant frequencies. However, the contribution of the frequency components higher than F2 to the articulatory control under the perturbations of F1 and F2 has not yet been investigated. In this study, a formant-AAF experiment was conducted in which a low-pass filter was applied to speech. The experimental results showed that the deviation in the compensatory response was significantly larger when a low-pass filter with a cutoff frequency of 3 kHz was used compared to that when cutoff frequencies of 4 and 8 kHz were used. It was also found that the deviation in the 3-kHz condition correlated with the fundamental frequency and spectral tilt of the produced speech. Additional simulation results using a neurocomputational model of speech production (SimpleDIVA model) and the experimental data showed that the feedforward learning rate increased as the cutoff frequency decreased. These results suggest that high-frequency components of the auditory feedback would be involved in the determination of corrective motor commands from auditory errors.

Effects of Vibrotactile Stimuli on Perception of Voiced and Unvoiced Bilabial Stop Consonants in Noise

A. Ono, M. Nakatani, A. Nakane, J. Watanabe, and S. Hiroya

Proc. of the 12th International Seminar on Speech Production (ISSP2020), pp. 194–197, July 2021.

Vibrotactile stimulation replicating laryngeal vibration has been reported to improve discrimination between degraded voiced and unvoiced consonants in consonant-vowel syllables. In this study, we investigated (1) whether or not vibrotactile stimulation in the consonant region biases the perception of unvoiced consonants toward voiced ones and (2) the relationship between the effect and auditory efficacy. Our results indicated that vibrotactile stimulation across the unvoiced consonant and vowel region, not just in the unvoiced consonant region, biased the consonant perception toward voiced consonants. Also, we found a non-linear effect of vibrotactile stimulation across the consonant-vowel region on auditory efficacy. Our findings will likely contribute to understanding how to improve intelligibility under noisy conditions and will likely help people with hearing impairments.