Shift to New Dimensions—Further Initiatives to Deepen Communication Science

Takeshi Yamada

Abstract

NTT Communication Science Laboratories aims to realize communication that *reaches the heart*, from person to person, and between people and computers. We are building fundamental theories in pursuit of the essence of people and of information, and working to create core technologies that will transform society. This article introduces some of our initiatives to push deeper in communication science, in areas including speech and audio processing, dialogue processing, human information science, sports brain science, and machine learning and optimization.

Keywords: artificial intelligence, communication science, corevo

1. Introduction

There have been some remarkable developments in artificial intelligence (AI) recently. Developments in deep learning, in particular, have achieved capabilities approaching those of humans in areas such as speech and image recognition and natural language processing, which were once considered human strengths that could not be matched by computers. At the NTT laboratories, we consider it important and necessary to use these leading-edge technologies and apply them to the issues we are facing. However, as these technologies spread, we must further strive to open new and next-generation areas that are not simple extensions of earlier work, and to make bold changes in our research themes.

We at NTT Communication Science Laboratories (CS Labs) aim to realize communication that *reaches the heart*, from person to person, and between people and computers. We are building fundamental theories in pursuit of the essence of people and of information, and working to create core technologies that will transform society. Our fields of research support the four types of AI that comprise the NTT Group's AI technology called corevo®. These are Agent-AI, Heart-Touching-AI, Ambient-AI, and Network-AI [1]. As such, we have focused mainly on media processing, human information science, and data and machine learning, and have also focused recently on sports brain science (**Fig. 1**). The Feature Articles in this issue introduce some of our initiatives to push deeper in these areas of communication science.

2. Speech and audio processing approaching human capabilities

Agent-AI, a component of corevo, supports interaction between humans and computers. CS Labs is working on speech and audio processing, image recognition, and natural language processing to provide a platform for Agent-AI. Speech recognition under conditions where a single person is speaking into a close-talking microphone has already matured to a level where it is used in everyday life, due to the rapid spread of smartphones and, more recently, devices such as AI speakers.

Research and development (R&D) trends are now shifting toward speech recognition with multiple people speaking freely around a table, some distance from the microphone. This requires increased performance of speech recognition itself, but it is also important to combine it with speech enhancement

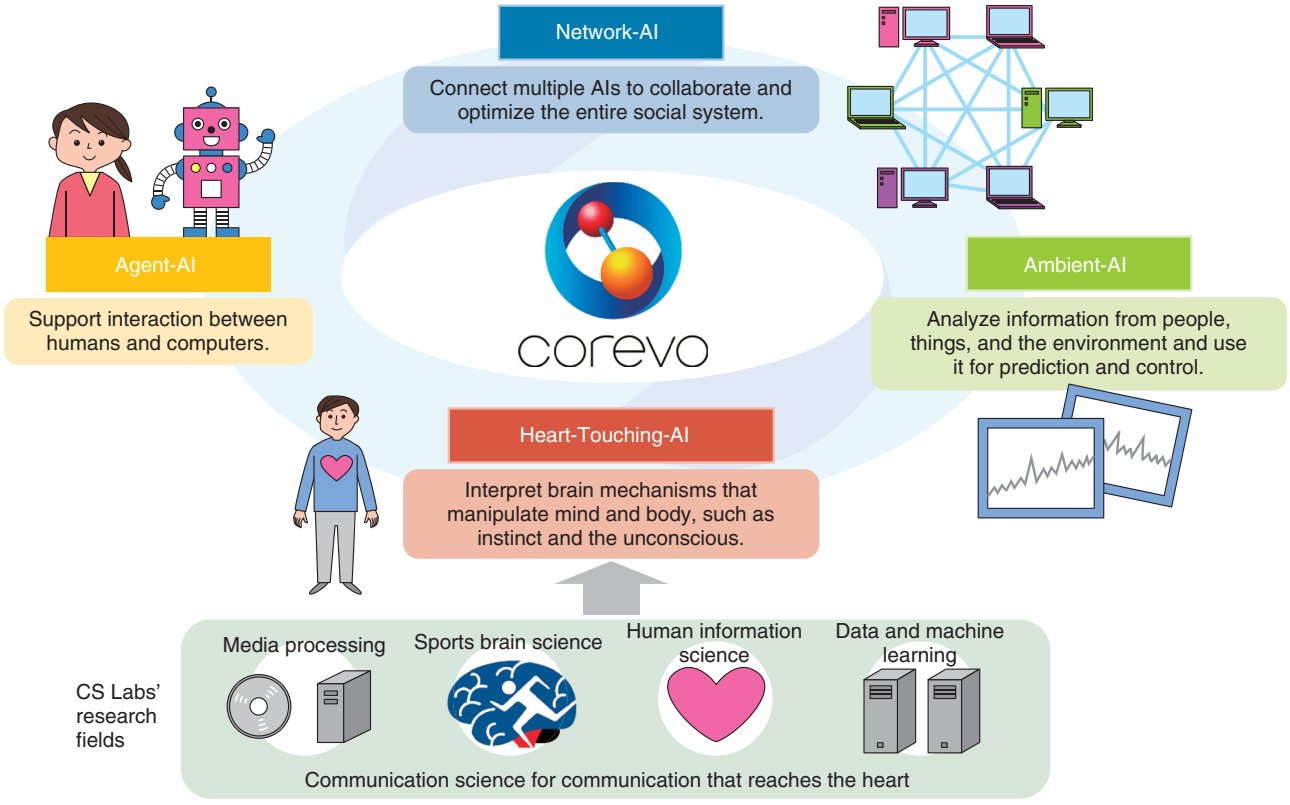


Fig. 1. NTT's four AI directions and communication science.

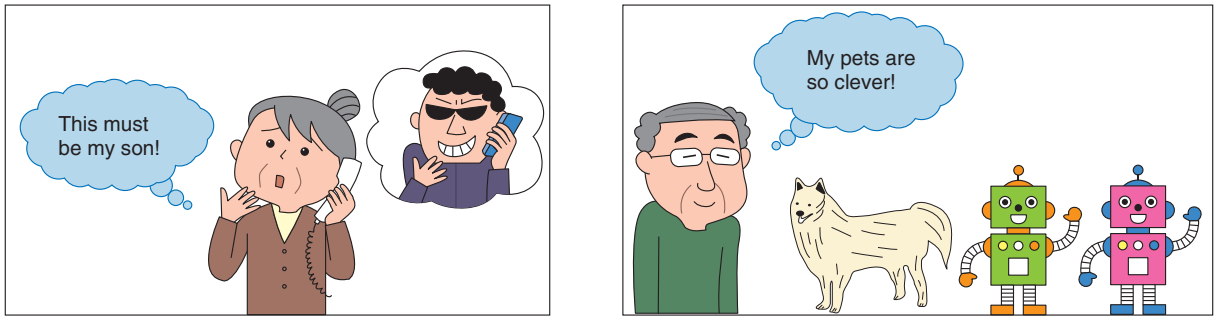
technologies such as denoising to remove background noise in noisy environments, and dereverberation to remove reverberation from the walls and floor of a room. A speech recognition technology from CS Labs combining these capabilities was entered in the 3rd CHiME Speech Separation and Recognition Challenge (CHiME-3) held in 2015, where it placed first among 25 participating organizations [2]. We are working to use these technologies for automatically creating the minutes of a meeting involving several people, while it is still in an experimental stage.

Humans can pick out the voice of a person they want to hear and can understand what that person is saying, in a discussion in a meeting or in a conversation during a party, when there are several people talking or when music is playing in the background. This is called selective listening. In the article in this issue entitled "SpeakerBeam: A New Deep Learning Technology for Extracting Speech of a Target Speaker Based on the Speaker's Voice Characteristics " [3], we introduce a technology that uses deep learning to implement selective listening using computers.

People are also able to visualize a scene in their mind just by listening to a sound. Cross-media scene analysis implements this type of behavior in computers. Deep learning can be applied to various media in a unified framework, so cross-media scene analysis based on deep learning transcends single-media processing, combines multiple types of media, and handles them in a complementary way. Cross-media scene analysis is introduced in the article, "Cross-media Scene Analysis: Estimating Objects' Visuals Only from Audio" [4].

3. Elimination of the human-AI gap

The capabilities of computers are approaching those of humans in certain situations such as those described above, but it will take more time for AI performance to advance beyond the complexity of the human brain. Nevertheless, humans are sometimes easily fooled by telephone payment scams and the like. This is related to certain human cognitive biases. For example, people tend to look only for evidence suited to their existing beliefs, hopes, and suppositions and to interpret circumstances accordingly. This



Confirmation bias: people tend to look only for evidence suited to their existing beliefs, hopes, and suppositions and to interpret circumstances accordingly

ELIZA effect: people have a tendency to arbitrarily anthropomorphize and sense intelligence from behavior that superficially appears to be intelligent

Fig. 2. Examples of cognitive biases.

type of cognitive bias is called a confirmation bias [5], which explains why elderly people easily assume a phone call asking for money is from their child and do not notice any inconsistencies.

People also have a well-known tendency to arbitrarily anthropomorphize and sense intelligence from behavior that superficially appears to be intelligent. This is another cognitive bias called the ELIZA effect, after a computer program developed at MIT (Massachusetts Institute of Technology) in the 1960s [6]. For example, people tend to think their pets are smarter than they really are (**Fig. 2**). Moreover, there are well-known examples of illusions such as the checker shadow illusion, which effectively demonstrates the fact that humans do not see physical quantities as they are, literally, before their very eyes [7].

Effective use of such cognitive biases and illusions in interfaces and for feedback is the key to filling the gap between humans, who are complex and imperfect, and AI, which is currently still immature and limited. Regarding cognitive biases, CS Labs is conducting research on dialogue processing for human-robot interaction, extending the level from casual conversation to serious debate, and at the same time, moving from dialogue with a single robot to that with multiple (two) robots. For conversation with a single person, a single robot may seem to be sufficient as a counterpart. However, multiple robots can be made to appear more intelligent to the human by dividing roles suitably and by utilizing robot-robot interaction that can take advantage of human cognitive biases more effectively. Dialogue with multiple robots seems more natural to the human and can be maintained for longer periods than that with a single robot, even if there are speech recognition mistakes or the dialogue context is lost [8].

CS Labs has also produced interfaces that use illusion, such as Buru-Navi, a device that produces an illusion of being pulled, and Hengento Projection, which produces an illusion of dynamic motion by projecting light onto a printed picture or photograph. The article in this issue, "Ukuzo—A Projection Mapping Technique to Give Illusory Depth Impressions to Two-dimensional Real Objects" [9], introduces a new technique that exploits human visual illusion.

4. Explaining implicit brain activities

To realize communication that reaches the heart, CS Labs is focusing on explaining implicit brain activity related to the basic human senses of vision, hearing, and motion. Research on interfaces as described above, using illusion, is also derived from this approach. This basic research on human information science is the Heart-Touching-AI component of corevo, an AI platform to interpret brain mechanisms that manipulate mind and body, such as instinct and the unconscious, and thereby support humans. Realizing communication that reaches the heart will surely be possible by developing AI that *touches the heart*, namely, Heart-Touching-AI.

Initiatives in sports brain science have recently become a new theme in Heart-Touching-AI [10]. This research makes use of knowledge from brain science that explains implicit brain activity, advanced information and communication technology such as wearable sensors and virtual reality, and machine learning to train the brain to win sports games. This research

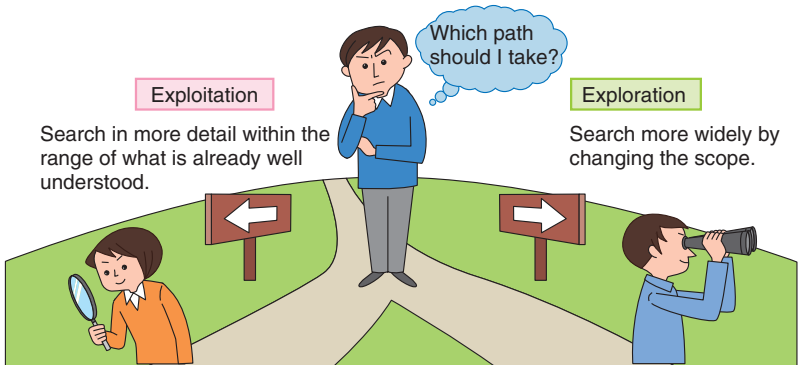


Fig. 3. Exploration-Exploitation Dilemma.

is not about making the body stronger but is focused on finding a way to coordinate the body parts optimally, and how to control the human mental state well. For example, on the basis of the fact that hearing has better time resolution than vision, we have proposed a method that provides intuitive feedback by measuring the activity of muscles in certain parts of the body and converting them to sounds in real time.

Beyond sports, it is also important to support people in their daily lives in order to maximize use of the implicit capabilities of their minds and bodies. The article in this issue, "Measuring, Understanding, and Cultivating Wellbeing in the Age of Technology" [11], introduces our challenge to identify design guidelines for human wellbeing, which cannot be grasped qualitatively at a glance, treating it quantitatively from a human-scientific perspective, and improving it.

5. Machine learning and optimization

With machine learning technology, it is possible from large amounts of data to automatically discover patterns that even human experts would not notice. The collective behavior of people is complex, as with a crowd of people walking in a town, but there are definite patterns, as each person has his or her own typical intention. At NTT, we are conducting R&D on technology that can predict the risks of congestion or delays in the near future from real-time data obtained by observing the flow of people. It can then automatically provide near-optimal on-line navigation to avoid such risks as efficiently as possible without impeding people in fulfilling their intentions. Here, we use techniques such as multi-agent simulation and Bayesian optimization [12].

With limited computing resources, it is often difficult to search for the optimal solution with a fine-toothed comb (i.e., a simple exhaustive search) when there are vast numbers of combinations. In such cases, we are obliged to choose whether to prioritize an Exploitation approach, in which we search in more detail within the range of what is already well understood, or an Exploration approach, in which we change the scope to search more widely. This is called the Exploration-Exploitation Dilemma (**Fig. 3**). Bayesian optimization is a method for efficiently narrowing down candidate solutions in a search, with consideration for the balance between Exploitation and Exploration [13]. NTT is conducting this R&D as a component of Ambient-AI, which provides AI as intelligence in the Internet of Things, analyzing information from people, things, and the environment, and using it for prediction and control.

On the other hand, there are also cases when data structures can be devised so that all combinations can be enumerated efficiently and a strictly optimal solution computed, even when a simple exhaustive search is difficult. The article "Network Reliability Optimization by Using Binary Decision Diagrams" [14] in this issue introduces examples of formerly unsolvable large-scale problems that were solved with surprising efficiency by using data structures such as binary decision diagrams.

As Ambient-AI progresses and AI technologies are used more often in networks, multiple AIs will collaborate and optimize overall social systems, forming Network-AI.

6. Future prospects

This article has introduced some new initiatives in communication science at CS Labs. New initiatives

also have accompanying risks. We cannot always obtain the scientific results we hope for immediately. In R&D as well, we face the Exploration-Exploitation Dilemma. We need to decide whether we should take an Exploitation approach, using the latest technologies and attempting to apply them skillfully to the problems we are facing, or give priority to Exploration, working to open up new dimensions that are not extensions of earlier work [15]. We at CS Labs will continue to face challenging problems, giving priority to Exploration approaches. To be sure, we need to be careful of cognitive biases in conducting research, so that we do not only look for and interpret evidence that is well suited to our own assumptions [16].

References

[1] T. Yamada, S. Takahashi, F. Naya, T. Ikebe, and S. Furukawa, "Artificial Intelligence Research Activities and Directions in the NTT Group," NTT Technical Review, Vol. 14, No. 5, 2016.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201605fa1.html

[2] NTT press release, "NTT Achieved Top Performance in a Noisy Speech Recognition International Challenge," Dec. 14, 2015.
http://www.ntt.co.jp/news2015/1512e/151214a.html

[3] 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, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa2.html

[4] G. Irie, H. Kameoka, A. Kimura, K. Hiramatsu, and K. Kashino, "Cross-media Scene Analysis: Estimating Objects' Visuals Only from Audio," NTT Technical Review, Vol. 16, No. 11, pp. 35–40, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa5.html

[5] M. Akiyama, "Psychology of Deception and Trouble for Seniors," Kokumin Seikatsu, Vol. 13, pp. 1–4, 2013 (in Japanese).

[6] J. Weizenbaum, "ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine," Communications of the ACM, Vol. 9, No. 1, pp. 36–45, 1966.

[7] Website of Illusion Forum by CS Labs, Checker Shadow (in Japanese).
http://www.kecl.ntt.co.jp/IllusionForum/v/checkerShadow/ja/index.html

[8] NTT press release issued on January 1, 2018 (in Japanese).
http://www.ntt.co.jp/news2018/1801/180131b.html

[9] T. Kawabe, "Ukuzo—A Projection Mapping Technique to Give Illusory Depth Impressions to Two-dimensional Real Objects," NTT Technical Review, Vol. 16, No. 11, pp. 30–34, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa4.html

[10] M. Kashino, "Understanding and Shaping the Athlete's Brain—NTT Sports Brain Science Project—," NTT Technical Review, Vol. 16, No. 3, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201803fa1.html

[11] J. Watanabe, Y. Ooishi, S. Kumano, M. Perusquía-Hernández, T. G. Sato, A. Murata, and R. Mugitani, "Measuring, Understanding, and Cultivating Wellbeing in the Age of Technology," NTT Technical Review, Vol. 16, No. 11, pp. 41–44, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa6.html

[12] F. Naya, M. Miyamoto, and N. Ueda, "Optimal Crowd Navigation via Spatio-temporal Multidimensional Collective Data Analysis," NTT Technical Review, Vol. 15, No. 9, 2017.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201709fa5.html

[13] E. Brochu, V. M. Cora, and N. de Freitas, "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning," arXiv:1012.2599, 2010.

[14] M. Nishino, T. Inoue, N. Yasuda, S. Minato, and M. Nagata, "Network Reliability Optimization by Using Binary Decision Diagrams," NTT Technical Review, Vol. 16, No. 11, pp. 25–29, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa3.html

[15] J. G. March, "Exploration and Exploitation in Organizational Learning," Organization Science, Vol. 2, No. 1, pp. 71–87, 1991.

[16] R. Nuzzo, "How Scientists Fool Themselves – and How They Can Stop," Nature, Vol. 526, No. 7572, pp. 182–185, 2015.

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 a 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 an 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 and a 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.