Technologies for Achieving Another Me

Atsushi Ohtsuka, Chihiro Takayama, Fumio Nihei, Ryo Ishii, and Toru Nishimura

Abstract

"Another Me," a grand challenge of Digital Twin Computing announced by NTT in 2020, aims to achieve the existence as a person's alter ego that appears to have the same intelligence and personality as an actual human and is recognized and active in society as that person. As the first step in achieving this, we constructed a digital twin that acts like the represented person and is capable of asking questions in line with the viewpoints held by that person. This article describes in detail the main technologies behind Another Me, namely, question-generation technology, body-motion-generation technology, and dialogue-video-summarization technology.

Keywords: Digital Twin Computing, grand challenges, Another Me

1. Introduction

The loss of various opportunities in one's life is becoming a social problem, such as the inability to participate in society despite an interest or desire to do so, possibly due to the difficulty of balancing childcare or caregiving of elderly family members with work. To dramatically increase the opportunities for a person to become more active and grow while expanding and merging the range of activities from the real world to the virtual world, we are taking up the challenge of achieving Another Me as a digital version of oneself [1, 2] (**Fig. 1**). Another Me will be able to participate actively as oneself in a way that transcends the limitations of the real world and share the results of such activities as experiences of oneself. There are three key technical issues surrounding this challenge: the ability to think and act autonomously as a human, the ability to have a personality like the represented person, and the ability to give feedback on the experiences obtained by Another Me. In this article, we describe in detail the main technologies for addressing these issues: question-generation technology, body-motion-generation technology, and dialogue-video-summarization technology.

2. Question-generation technology

To get Another Me to act autonomously, the digital twin must be able to make decisions about its next action. A means of collecting information as decision-making material is essential for this to be possible. We developed question-generation technology focusing on questions as a means of collecting information. Given input in the form of documents or text of conversations, this technology can automatically generate questions evoked by that input text. Generating questions and obtaining the answers to them in this manner enables a digital twin to autonomously collect information that it lacks.

Question-generation technology differs from current question-generation technology in two ways. First, the content of questions to be generated can be controlled by *viewpoint*. A person's values and position in life are greatly reflected by the questions that the person asks. Taking as an example an internal memo circulated within a company for decision-making purposes, we can expect the sales department to ask many questions about prices and costs, while a review by the legal department would likely include many questions related to legal compliance. By inputting a viewpoint label simultaneously with text,

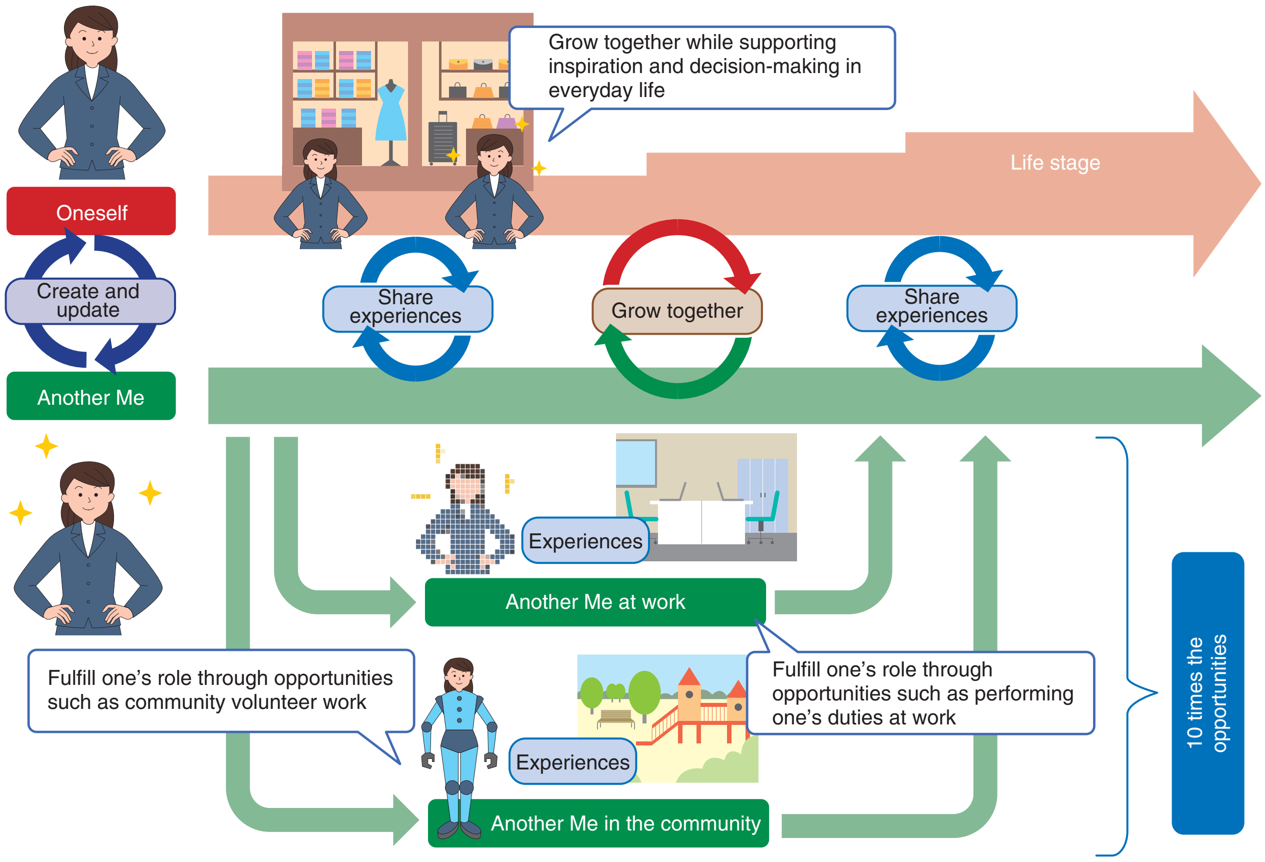
Oneself
Create and update
Another Me
Grow together while supporting inspiration and decision-making in everyday life
Life stage
Share experiences
Grow together
Share experiences
Experiences
Another Me at work
Fulfill one's role through opportunities such as community volunteer work
Fulfill one's role through opportunities such as performing one's duties at work
Experiences
Another Me in the community
10 times the opportunities

Fig. 1. Another Me.

the question-generation technology can generate questions corresponding to the input viewpoint. For example, inputting "money" as a viewpoint label will generate questions related to an amount of money as in costs, while inputting "law" as a viewpoint label will generate questions related to regulations and compliance. The viewpoint label to be input can be changed depending on the values, affiliated organization, etc. of a digital twin incorporating question-generation technology. This means that the digital twin can ask optimal questions according to its current thoughts and situation enabling it to collect the information needed to make decisions.

The question-generation technology can also recognize the input viewpoint and text content to automatically determine whether to generate questions. If content related to the input viewpoint has already been entered, or if an answer can be understood by simply reading the input text, there is a function for blocking the generation of questions. For example, when inputting the viewpoint label "money," questions related to money will be generated if there are no entries for price, cost, etc., but if there is a sufficient number of entries related to money such as price and cost, no questions will be generated. Therefore, when a digital twin is making decisions, it will ask questions only when having an insufficient amount of information and will quit asking questions and move on to processing for the next decision once it has collected a sufficient amount of information.

3. Body-motion-generation technology

To obtain the feeling that Another Me has the same personality as an actual person, its appearance is, of course, important, but making voice, way of speaking, and body motion like that person is also important. To date, we have clarified that differences in body motion, such as facial expressions, face and eye movements, body language, and hand gestures, can be felt as differences in personality traits [3] and that those differences can be a significant clue to

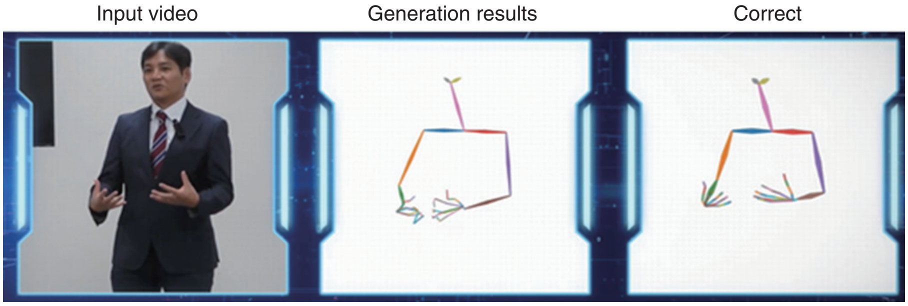

Fig. 2. Example input video of target person, results of generating body motion, and actual (correct) body motion from the input video.

distinguishing individuals [4].

Giving such body motions to an autonomous system in the manner of Another Me (for example, an interactive agent) and manipulating them accordingly is a difficult technical problem from an engineering standpoint. This problem has so far been addressed in technologies for generating human-like body motion and body motion corresponding to personality traits from utterance text [5, 6]. However, the generation of the same movements as those of a specific, actual person has not yet been achieved.

Against this background, we developed technology for automatically generating body motion like that of a real person on the basis of uttered-speech information in the Japanese language. We construct a model for automatically generating body motion like that of an actual person by preparing only video data (time-series data consisting of speech and images) of that person. We then use this model to automatically generate movements like those that person would make when speaking by simply inputting uttered-speech information. The first step for constructing the generation model involves using automatic speech-recognition (ASR) technology to extract utterance text from the voice data included in the video data of the target person when speaking. The positions of joint points on the person's body are also extracted automatically from the video data. The next step involves training the generation model by using a deep learning method called a generative adversarial network that can generate the positions of body joint points from voice and utterance text. To construct a model during this learning process that can capture even a person's subtle habits and generate a wide range of movements, we devised a mechanism for achieving high-level resampling of the data during training. As a result, we have achieved the highest level of performance in subjective evaluations with regard to a person's likeness and naturalness (as of October 2021) [7]. With this technology, we constructed a model for generating body motion using speech in the Japanese language as input. **Figure 2** shows an example input video of the target person, results of generating body motion, and actual (correct) body motion from the input video.

This body-motion-generation technology can automatically generate body motion of a particular person in Another Me, computer-graphics characters, humanoid robots, etc. It can also easily generate avatar body motions of a particular person in a web meeting from only uttered-speech information.

We plan to construct a model that can train a body-motion-generation model from a small amount of data and develop generation technology that can achieve a better likeness of an existing person.

4. Dialogue-video-summarization technology

Dialogue-video-summarization technology summarizes recorded dialogue over a length of time shorter than the actual meeting and generates video not only of the content of that dialogue but also of the atmosphere of the place where the dialogue occurred.

Our aim is to achieve a society in which humans and Another Me can grow together. It will not be sufficient to simply use Another Me as a surrogate for oneself—it will also be necessary to efficiently feed the experiences obtained by Another Me back to oneself. It will also be necessary to convey not only the content of what Another Me does but also the emotions likely to be felt at that time and at that place so that the represented person can treat what Another Me does as one's actions. We are researching dialogue-video-summarization technology targeting dialogue as one technology for feeding back the experiences of Another Me to the represented person.

As the first step in this research, we are taking up dialogue-condition-estimation technology and video-summarizing technology to enable efficient reviews of meetings. These combined technologies analyze and reconfigure the video of a meeting taken with a small camera or obtained from a web meeting and generate a video summary.

(1) Dialogue-condition-estimation technology

This technology uses a collection of information in various forms (multimodal information) such as the speech and behavior of the speaker including video of the dialogue as a clue to estimating diverse dialogue conditions. These include the importance and persuasiveness of each utterance, the intention and motivation of utterances, the personality traits and skills of individual participants, and the role of participants within the conversation [8–13]. The technology constructs a highly accurate estimation model by using a variety of machine-learning techniques. These include a multimodal deep-learning technique, which collects a variety of information such as the temporal change in the behavior of each participant, synchronization of movements among participants, change in the speaker's voice, and content of utterances for learning purposes, and a multitask-learning technique that simultaneously estimates multiple dialogue conditions.

(2) Video-summarizing technology

This technology uses the results of estimating various dialogue conditions obtained with the technology described above to extract important comments, questions asked of other participants, and comments made in reaction to other participants' opinions and reconfigure the meeting video to a shorter video of about one fourth the actual time. Subtle nuances of a participant's comments can be conveyed by the facial expressions or the tone of voice of that participant included in the summarizing video.

This dialogue-video-summarization technology enables a user to efficiently grasp in a short period the flow of a discussion, state of participants (e.g., an approving or opposing attitude to an opinion), and atmosphere (energy level) of the meeting that one was unable to attend. Our future goal is to summarize and convey not only dialogue among humans but also between a human and digital twin and even among digital twins. In the area of digital-twin behavior beyond dialogue, we aim to research techniques for feeding back information to humans in a more efficient manner with a higher sense of presence.

References

[1] Press release issued by NTT, "NTT Announces New R&D Projects of Digital Twin Computing," Nov. 13, 2020. https://group.ntt/en/newsrelease/2020/11/13/201113c.html

[2] R. Kitahara, T. Kurahashi, T. Nishimura, I. Naito, D. Tokunaga, and K. Mori, "Research and Development of Digital Twin Computing for Creating a Digitalized World," NTT Technical Review, Vol. 19, No. 12, pp. 16–22, 2021. https://doi.org/10.53829/ntr202112fa1

[3] Y. Nakano, M. Ooyama, F. Nihei, R. Higashinaka, and R. Ishii, "Generating Agent's Gestures that Express Personality Traits," The Transactions of Human Interface Society, Vol. 23, No. 2, pp. 143–154, 2021 (in Japanese).

[4] C. Takayama, M. Goto, S. Eitoku, R. Ishii, H. Noto, S. Ozawa, and T. Nakamura, "How People Distinguish Individuals from Their Movements: Toward the Realization of Personalized Agents," Proc. of the 9th International Conference on Human-Agent Interaction (HAI 2021), pp. 66–74, Nov. 2021.

[5] R. Ishii, R. Higashinaka, K. Mitsuda, T. Katayama, M. Mizukami, J. Tomita, H. Kawabata, E. Yamaguchi, N. Adachi, and Y. Aono, "Methods of Efficiently Constructing Text-dialogue-agent System using Existing Anime Characters," Journal of Information Processing, Vol. 29, pp. 30–44, Jan. 2021.

[6] R. Ishii, C. Ahuja, Y. Nakano, and L. P. Morency, "Impact of Personality on Nonverbal Behavior Generation," Proc. of the 20th ACM International Conference on Intelligent Virtual Agents (IVA 2020), Article no. 29, pp.1–8, Oct. 2020.

[7] C. Ahuja, D. W. Lee, R. Ishii, and L. P. Morency, "No Gestures Left Behind: Learning Relationships between Spoken Language and Freeform Gestures," Proc. of Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1884–1895, Nov. 2020.

[8] F. Nihei and Y. Nakano, "Validating the Effectiveness of a Meeting Summarizing Browser that Implements an Important Utterance Detection Model Based on Multimodal Information," The Transactions of Human Interface Society, Vol. 22, No. 2, pp. 137–150, 2020 (in Japanese).

[9] R. Ishii, K. Otsuka, S. Kumano, R. Higashinaka, and Y. Aono, "Estimation of Personal Empathy Skill Level Using Dialogue Act and Eye-gaze during Turn-keeping/changing," IPSJ Journal, Vol. 62, No. 1, pp. 100–114, 2021 (in Japanese).

[10] R. Ishii, S. Kumano, and K. Otsuka, "Estimating Empathy Skill Level Using Gaze Behavior Depending on Participant-role during Turn-changing/keeping," The Transactions of Human Interface Society, Vol. 20, No. 4, pp. 447–456, 2018 (in Japanese).

[11] T. Onishi, A. Yamauchi, A. Ogushi, R. Ishii, Y. Aono, and A. Miyata, "Analyzing Head and Face Behaviors along with Praising," IPSJ Journal, Vol. 62, No. 9, pp. 1620–1628, 2021 (in Japanese).

[12] R. Ishii, X. Ren, M. Muszynski, and L. P. Morency, "Multimodal and Multitask Approach to Listener's Backchannel Prediction: Can Prediction of Turn-changing and Turn-management Willingness Improve Backchannel Modeling?", Proc. of IVA 2021, pp. 131–138, Sept. 2021.

[13] R. Ishii, X. Ren, M. Muszynski, and L. P. Morency, "Can Prediction of Turn-management Willingness Improve Turn-changing Modeling?", Proc. of IVA 2020, No. 28, pp. 1–8, Oct. 2020.

Atsushi Ohtsuka

Research Engineer, NTT Digital Twin Computing Research Center.

He received an M.S. in informatics from the University of Tsukuba, Ibaraki, and joined NTT in 2013. He received a Ph.D. in informatics from the University of Tsukuba in 2020. His research interests are natural language processing and information retrieval. He has been in charge of research for Digital Twin Computing at NTT Digital Twin Computing Research Center.

Chihiro Takayama

Research Engineer, NTT Digital Twin Computing Research Center.

He received an M.S. in information engineering and M.A. in business from Waseda University, Tokyo, and joined NTT in 2009. His research interests include human-computer interaction. Since 2020, he has been in charge of research for Digital Twin Computing at NTT Digital Twin Computing Research Center.

Fumio Nihei

Researcher, NTT Digital Twin Computing Research Center.

He received a Ph.D. in science and technology from Seikei University, Tokyo and entered the university as a postdoc researcher in 2019. After working as a postdoc, he joined NTT in 2021. His research interests are multimodal interaction and social signal processing. Since 2021, he has been in charge of research for Digital Twin Computing at NTT Digital Twin Computing Research Center.

Ryo Ishii

Distinguished Researcher, NTT Digital Twin Computing Research Center.

He received an M.S. in engineering from the Tokyo University of Agriculture and Technology in 2008 and joined NTT the same year. He received a Ph.D. in informatics from Kyoto University in 2013. He was a visiting scholar at Carnegie Mellon University from 2019 to 2020. He is currently a distinguished researcher at NTT Human Informatics Laboratories. His research interests are multimodal interaction and social signal processing. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Japanese Society for Artificial Intelligence (JSAI), and Human Interface Society.

Toru Nishimura

Senior Research Engineer, NTT Digital Twin Computing Research Center.

He received an M.S. in information engineering from Nagoya University in 1998 and joined NTT the same year. He has been involved in the research and development of an operation support system for IT systems including networks. Since 2020, he has been in charge of research for Digital Twin Computing at NTT Digital Twin Computing Research Center.