Creating "Shido Twin" by Using Another Me Technology

Atsushi Fukayama, Ryo Ishii, Akira Morikawa, Hajime Noto, Shin-ichiro Eitoku, Yusuke Ijima, and Hiroki Kanagawa

Abstract

"Cho Kabuki 2022 Powered by NTT," a kabuki play sponsored by Shochiku Co., Ltd., is the first social implementation of Another Me, a technology for creating a human digital twin that reproduces the appearance and internal aspects of a real person while acting autonomously. We created a digital twin of the star of Cho Kabuki, Shido Nakamura, and call it "Shido Twin," which performed in the play alongside virtual diva Hatsune Miku. This article overviews this initiative and describes the main technologies behind Cho Kabuki 2022, i.e., automatic body-motion generation and deep neural network-based text-to-speech synthesis.

Keywords: digital human, AI, kabuki

1. Introduction

Another Me, one of the grand challenges concerning Digital Twin Computing (DTC), aims to extend opportunities for self-realization and personal growth beyond constraints such as time, space, and handicaps by having a digital twin of a real person act in place of that person in society. We have set *identity*, *autonomy*, and *oneness* (**Fig. 1**) as requirements for creating one's Another Me and are researching and developing technologies to satisfy those requirements.

For a person's Another Me to act as that person in society, it must first have *identity*, which means that it is recognized as that person by reproducing their external characteristics, such as appearance and movements, as well as their internal aspects, such as personality and values. To overcome time, physical, and cognitive handicaps, one's Another Me must then have *autonomy* so that it can understand situations, make judgments, and act in the same way as the person it represents without that person having to operate or give instructions at every step. To acquire a sense of accomplishment through self-realization and personal growth from the results of the activities of one's Another Me that fulfill the first-two requirements, it is necessary to maintain *oneness* between the person and their Another Me by feeding back the results to the person with a real feeling as if that person had experienced them.

2. Initiative to create Shido Twin

An entity that completely satisfies all three requirements can be called Another Me; however, in reality, it is necessary to determine which requirements should be satisfied to what extent in accordance with the application area of that entity. Taking the first step in the social implementation of Another Me, we focused on creating identity and took on the challenge of recreating an actor on a theater stage as a venue for this creation. In cooperation with Shochiku Co., Ltd., which has been working on "Cho Kabuki"—combining kabuki (Japan's traditional theater) and NTT's latest technologies, i.e., automatic body-motion-generation and deep neural network (DNN) text-to-speech (TTS) synthesis, we created a digital twin (called Shido Twin) of the star of Cho Kabuki,

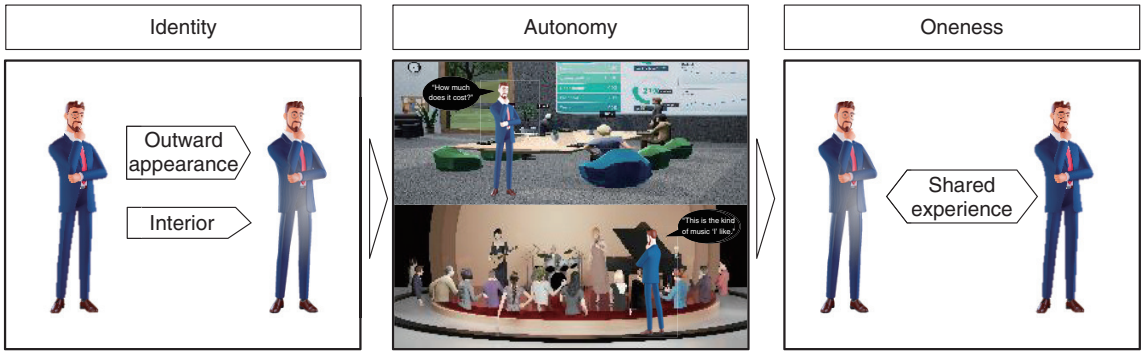


Fig. 1. Three requirements of Another Me.

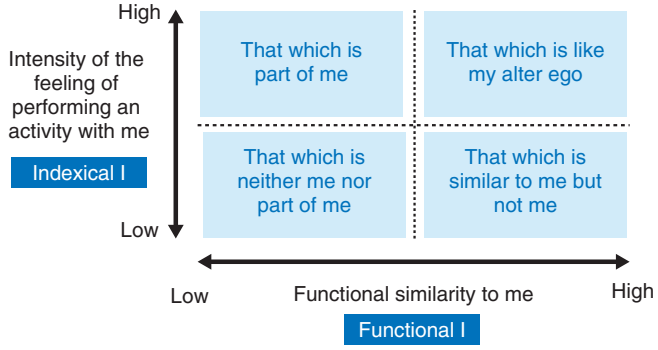


Fig. 2. The two-layered structure of the "I" concept.

Shido Nakamura, and had Shido Twin greet the audience in place of him. Since many in the audience of Cho Kabuki are fans of Shido Nakamura and Cho Kabuki, satisfying the demand of a high level of personal identity, especially in terms of external appearance, is challenging. That demand leads to the question of what does it mean to recognize the identity of an entity such as Another Me that is not the actual person? We have explored this question through co-creation with experts in philosophy and have come to understand identity along two axes: "Functional I" and "Indexical I" (**Fig. 2**).

Functional I refers to the fact that a person's external characteristics, such as appearance and movement, as well as skills and abilities, are the same as those of the person in question. In consideration of this fact, this project involved about half a day of studio recording to create an elaborate three-dimensional computer graphics model of Shido Nakamura and construct a machine-learning model that can generate the gestures and voice similar to the actor.

In contrast, Indexical I is the idea that Another Me can share *indexicality* (consciousness that points to the person such as "he," "she," and "I") by making the past experiences that characterize the person felt by Another Me. For the project, targeting fans of Cho Kabuki, we asked Shido Nakamura to perform the gestures and vocalizations that fans have come to know from past Cho Kabuki performances to reproduce the rousing of the audience. Costumes and dialogues that would not be out of place in a traditional cultural setting as well as the interactions with the live performers on stage were finalized after close consultation with Shochiku. The technologies for creating Shido Twin are described in the following sections.

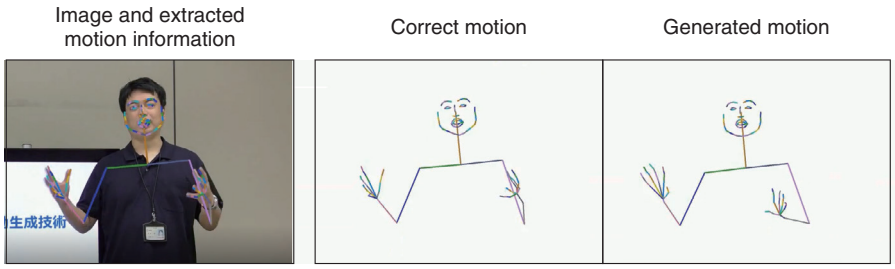


Fig. 3. Example of motion generated using automatic body-motion-generation technology.

3. Automatic body-motion-generation technology that can reproduce even the most subtle individual habits from a small amount of data

It is considered important for Another Me to be felt as having the same personality, voice, speech, and body motion of the real person, not to mention appearance. We have previously shown that differences in body motions, such as facial expressions, facial and eye movements, and body gestures, are major cues for perceiving differences in personality traits [1] and identifying others [2]. It is therefore important to control the motion of Another Me so that it automatically generates the body motion of the person in question, which is a difficult technical challenge from an engineering standpoint. We have been investigating technologies for generating human-like body motions and body motions based on personality traits from spoken text [3, 4]; however, we have not been able to generate motions that can mimic the subtle habits of a specific person in real life.

We developed a technology for automatically generating body motions, in a manner that mimics the subtle habits of a real person during speech, on the basis of Japanese speech and its textual information. Simply by preparing video data (time-series data of audio and body images) of a real person, it is possible to construct a generative model that automatically generates body motions that are typical of that person. By using this generative model, a user can automatically generate a person-specific behavior during speech by simply inputting speech sounds and their text information. First, speech-recognition technology is used to extract speech text from the speech data contained in the video of the target person, and the positions of joints of the body are automatically extracted from the image data. Next, a deep-learning generative model called a GAN (generative adversarial network), which can generate the positions of joints of the body from speech and speech text, is trained. To construct a model that can generate a wide range of motions by capturing even the most detailed habits of a person during training, we devised a mechanism for appropriately resampling the training data during training and have maintained the world's highest performance in subjective evaluation of human-like qualities and naturalness (as of November 2022) [5]. With this technology, we constructed a model for generating body motions by using Japanese speech as input. Examples of the input video of the person, result of body-motion generation, and actual correct body motion in the input video are shown in **Fig. 3**. We are also currently developing a learning method using the mechanism *few-shot learning* that can train models with only a small amount of data and without using a large amount of video data (training data) from a specific individual. With this method, we constructed a motion-generation model that can reproduce even the most subtle habits of Shido Nakamura from a small amount of video data (approximately 10 minutes) of him speaking and used the motion-generation results to control the motion of Shido Twin.

4. Saxe, a low-cost DNN TTS synthesis engine that reproduces a variety of speakers and tones

Voice is one of the most-important elements in reproducing a person's personality. TTS synthesis technology should be able to reproduce the desired speaker's voice with high accuracy. However, generating high-quality speech of a desired speaker requires a large amount of speech data uttered by that speaker, for example, up to 20 hours to produce high-quality synthesized speech with the concatenative TTS method Cralinet [6]. Consequently, the cost of recording voices and constructing databases has been

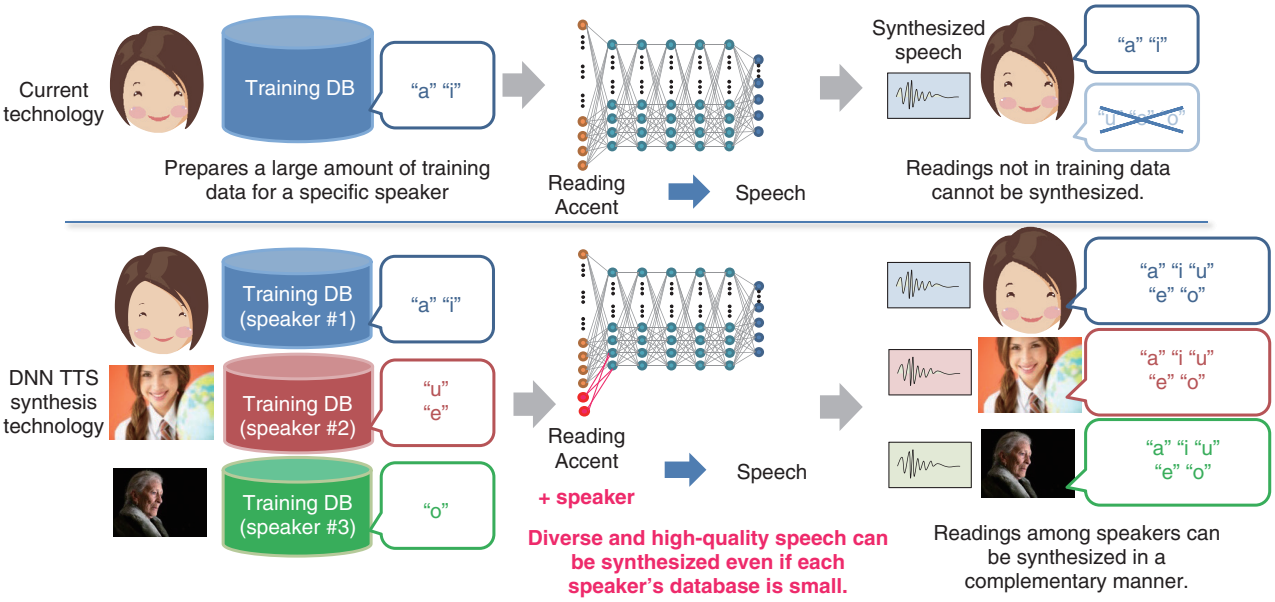


Fig. 4. DNN TTS synthesis technology for reproducing diverse speaker characteristics.

a major issue in regard to achieving TTS of desired speakers.

To address this issue, we developed a TTS engine called Saxe, which is based on a DNN [7]. Saxe uses a speech database (built from the utterances of a large number of speakers) and deep learning to synthesize high-quality speech of the desired speaker from a small amount of speech data. The characteristic feature of Saxe is that a large amount of speakers' speech data is modeled with a single DNN (**Fig. 4**). Information necessary for speech production, such as pronunciation and accent, is learned from a large amount of pre-prepared speech data, and the speaker characteristics of a desired speaker are learned from a small amount of speech data of the desired speaker. It is thus possible to generate high-quality speech even with a small amount of speech data of the desired speaker.

It is also important to reproduce the performance as well as the voice of the person by reproducing speech features, such as inflection and speech rhythm, as well as the tone of the voice. However, very specific articulations, such as those used in acting, make it very difficult to reproduce speech rhythms from a small amount of speech data. In response to this difficulty, we are developing a technology for extracting inflection and speech rhythm from a small amount of speech data of a desired speaker. When a small amount of speech is input to the DNN, the DNN outputs a low-dimensional vector of the inflection and speech rhythm of that speech [8]. During speech synthesis, the resulting low-dimensional vectors are combined with the aforementioned speech-synthesis DNN to generate synthesized speech with the desired tone, inflection, and speech rhythm of the speaker's voice.

5. Concluding remarks

Shido Twin created with these technologies performed in "How to Appreciate Cho Kabuki," an explanation of the appeal of Cho Kabuki, as one of the performances in "Cho Kabuki 2022 Powered by NTT," and it was well received (**Fig. 5**). This initiative showed that the Another Me technology can reproduce identity at a quality that can satisfy the demands of commercial performances. We will continue to develop the identity and autonomy of Shido Twin and strive to demonstrate the social value of Another Me in a variety of settings.

Shido Twin

Kuniya Sawamura, Shido Twin, and Choshi Nakamura

@Shochiku Cho Kabuki 2022 Powered by NTT
https://group.ntt/jp/newsrelease/2022/08/03/pdf/220803aa.pdf

Fig. 5. Scenes from the performance of Shido Twin.

References

[1] Y. Nakano, M. Ooyama, F. Nihei, R. Higashinaka, and R. Ishii, "Generating Agent's Gestures that Express Personality Traits," Transactions of the Human Interface Society, Vol. 23, No. 2, pp. 153–164, 2021.

[2] 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, Online, Nov. 2021.

[3] 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 Character," Journal of Information Processing, Vol. 29, pp. 30–44, Jan. 2021.

[4] R. Ishii, C. Ahuja, Y. I. 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, Online, Oct. 2020.

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

[6] K. Mano, H. Mizuno, H. Nakajima, N. Miyazaki, and A. Yoshida, "Cralinet—Text-To-Speech System Providing Natural Voice Responses to Customers," NTT Technical Review, Vol. 5, No. 1, pp. 28–33, Jan. 2007.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200701028.pdf

[7] Y. Ijima, N. Kobayashi, H. Yabushita, and T. Nakamura, "Saxe: Text-to-Speech Synthesis Engine Applicable to Diverse Use Cases," NTT Technical Review, Vol. 18, No. 12, pp. 48–52, Dec. 2020.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202012fa7.html

[8] K. Fujita, A. Ando, and Y. Ijima, "Phoneme Duration Modeling Using Speech Rhythm-Based Speaker Embeddings for Multi-Speaker Speech Synthesis," Proc. of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), pp. 3141–3145, Brno, Czech Republic, Aug./Sept. 2021.

Atsushi Fukayama
Senior Research Engineer, Group Leader, Another Me Research Group, NTT Digital Twin Computing Research Center.

He received an M.S. in informatics from Kyoto University in 1999 and joined NTT the same year. After working on media recognition technology, research and development of human-computer interaction technology, and practical application development of network services, he began leading the Another Me Research Group at the NTT Digital Twin Computing Research Center in 2021.

Ryo Ishii
Distinguished Researcher, NTT Digital Twin Computing Research Center.

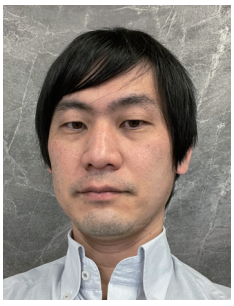
He received an M.S. in engineering from the Tokyo University of Agriculture and Technology and joined NTT in 2008. He received a Ph.D. in informatics from Kyoto University in 2013. His research interests are multimodal interaction and social signal processing. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Japanese Society for Artificial Intelligence (JSAI), and Human Interface Society.

Akira Morikawa
Research Engineer, NTT Digital Twin Computing Research Center.

He received an M.S. in engineering from Kobe University and joined NTT in 2011. His research interests are multimodal interaction and security.

Hajime Noto
Senior Research Engineer, Supervisor, NTT Digital Twin Computing Research Center.

He received a B.E. and M.E. in industrial engineering from Kansai University, Osaka, in 1997 and 1999. In 1999 he joined NTT Cyber Space Laboratories, where he researched three-dimensional input systems. His fields of interest are computer vision, video communication, and virtual reality. He is a member of the Institute of Image Information and Television Engineers and is currently involved in the research and development of human digitization.

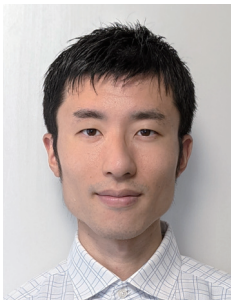


Shin-ichiro Eitoku
Senior Research Engineer, NTT Digital Twin Computing Research Center.

He received an M.E. and Ph.D. in information science and technology from the University of Tokyo in 2006 and 2013 and joined NTT in 2006. His research interests include information systems and multimedia systems for communications.

Yusuke Ijima
Distinguished Researcher, Human Insight Laboratory, NTT Human Informatics Laboratories.

He received a B.E. in electric and electronics engineering from National Institution for Academic Degrees and University Evaluation from Yatsushiro National College of Technology, Kumamoto, in 2007, and M.E. and Ph.D. in information processing from Tokyo Institute of Technology in 2009 and 2015. He joined NTT Cyber Space Laboratories in 2009, where he engaged in the research and development of speech synthesis. His research interests include speech synthesis, speech recognition, and speech analysis. He received the Awaya Prize Young Researcher Award of the Acoustical Society of Japan (ASJ) in 2018 and Maejima Hisoka Award, Encouragement Award in 2021. He is a member of ASJ, IEICE, and the International Speech Communication Association (ISCA).



Hiroki Kanagawa
Research Engineer, Human Insight Laboratory, NTT Human Informatics Laboratories.

He received an M.E. in information processing from Tokyo Institute of Technology in 2013. He was a research engineer at Information Technology R&D Center, Mitsubishi Electric Corporation from 2013 to 2017, where he engaged in the research and development of speech recognition, before joining NTT in 2017. His research interests are speech synthesis, voice conversion, and speech recognition. He is a member of ASJ and ISCA.