Media Processing Technology for Achieving Hospitality in Information Search

Kugatsu Sadamitsu, Jun Shimamura, Go Irie, Shuhei Tarashima, Taiga Yoshida, Ryuichiro Higashinaka, Hitoshi Nishikawa, Noboru Miyazaki, Yusuke Ijima, and Yukihiro Nakamura

Abstract

This article introduces services for achieving *hospitality* in information search activities. These services are designed to assist individual users in their surroundings during day-to-day activities. We also introduce subject identification technology based on images, natural language processing technology for understanding people and responding naturally, and speech synthesis technology capable of generating synthesized speech with a wide variety of speakers and speaking styles, all of which support these services.

Keywords: subject identification, natural language processing, speech synthesis

1. *Hospitality* in information search

At NTT, our aim is to implement services that provide support that is detailed and appropriate to the user context and that is intended for individual users in various everyday activity scenarios. We introduce specific examples of this below.

1.1 Service providing users with information about unknown entities

This service presents users with information about unknown items that they are curious about, such as unfamiliar cuisine or folk art, tailored to those users (**Fig. 1**). This service is based on subject recognition technology developed by NTT laboratories. It rapidly identifies items using the cameras in smartphones and tablets that we use every day. The service uses goods-related information available on the Internet and combines it with individual characteristics of each user's culture and tastes, to provide users with content that is suited to them and in their own language. We can anticipate various services such as those that present information about Japanese cuisine or folk art to foreigners visiting Japan in their native languages, those that display cooking ingredients to people who are careful about what they eat because of their cultural background or food allergies, and those that provide shoppers with word-of-mouth information relating to products.

1.2 Agent service with soft toys that understand user's intention

People enjoy spending time with family and friends, and there are often occasions during activities such as traveling and watching sports where someone wants to pull a smartphone out and check information. However, this immediately results in a sense of being isolated from the circle of family and friends, and the person is enclosed in his own world.

What if it were possible to comprehend a user's intention in the middle of a casual conversation between users and to convey the information that is

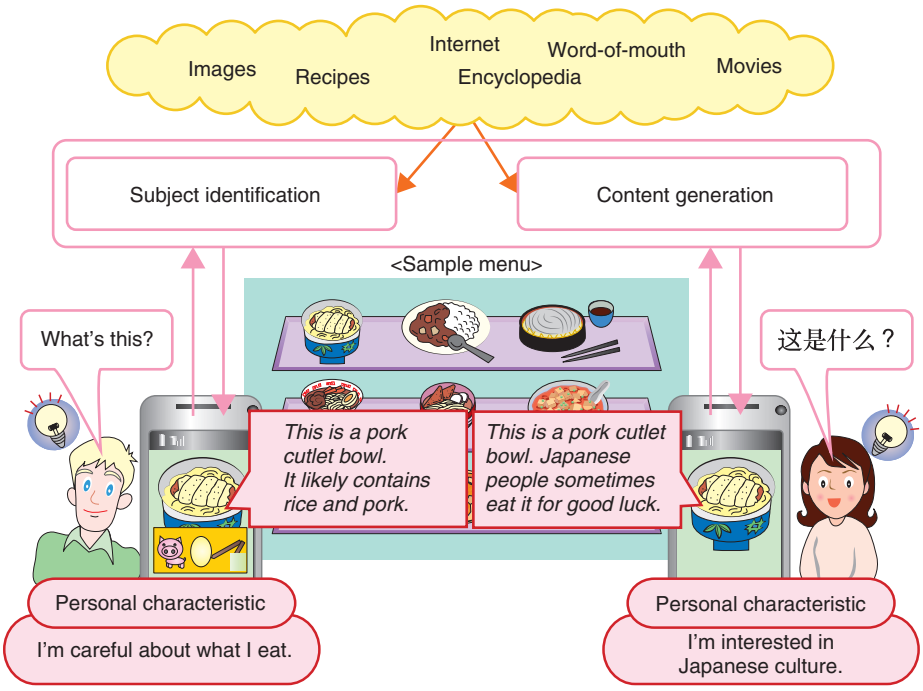

Fig. 1. Example of providing users with relevant information about unknown items.

required in a timely manner? It would seem as if a new member had joined that circle to provide the information to the user, without breaking up the circle of people or isolating any of the existing members.

The agents we have developed have a material form (such as a soft toy or puppet) and coexist physically within a circle of people, as represented by the teddy bear shown in **Fig. 2**. After comprehending the person's intent, the agent then generates utterances with the appropriate content and volume. They can behave naturally within the circle of people by speaking with a wide variety of synthesized speech. In the near future, we could have a world in which such agents are present within various circles of people.

2. Media processing technology that supports hospitality in information search

At NTT, we are working on the research and development (R&D) of subject identification technology based on images, natural language processing technology for understanding people and responding naturally, and speech synthesis technology capable of generating synthesized speech with a wide variety of speakers and speaking styles in order to implement services that add hospitality to information search.

2.1 Subject identification technology based on images

To get close to the user and provide information that depends on the user's situation, it is necessary to have a computer that can identify and comprehend the world and objects in the same way a person can. We introduce here our subject identification technology that identifies a photographed subject from an image captured by a camera that corresponds to a person's eyes. When subjects are identified from images, it is necessary to prepare reference images relating to those subjects beforehand. To cope with differences in photographic, environmental, and illumination conditions, however, it is usually necessary to prepare a large number of reference images for one subject, and such preparation requires a great deal of work. Here, we introduce our technologies for identifying subjects; one is capable of identifying three-dimensional (3D) subjects, and the other uses a cloud data application for subject identification. These technologies are expected to greatly reduce the work of preparing images in advance.

(1) 3D subject identification technology [1]

This technology makes it possible to identify even a 3D subject from a small number of reference images

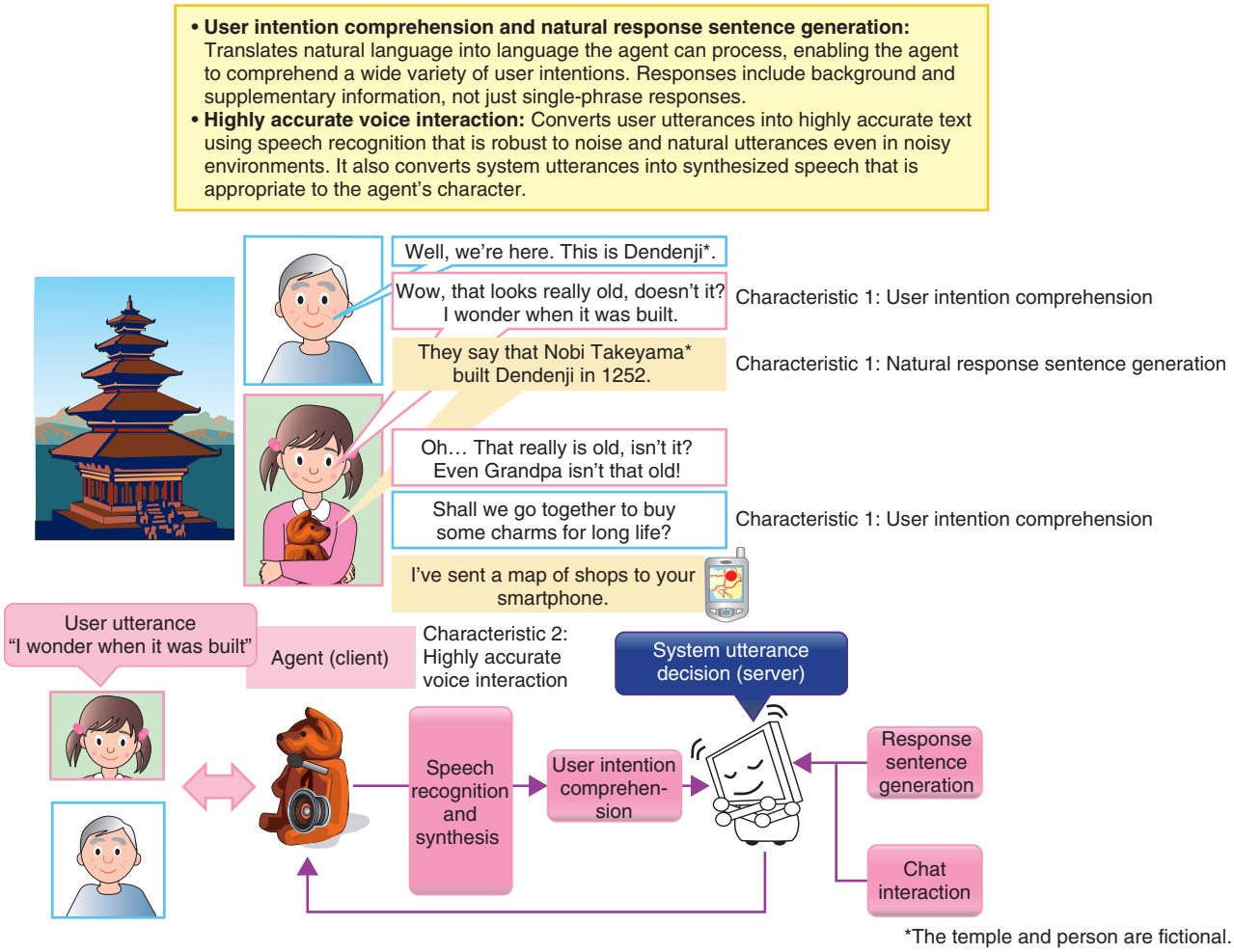

Fig. 2. Agent service using soft toys that understand user's intention.

with a high level of accuracy (**Fig. 3**). In contrast to flat objects such as books or compact discs, the appearance of a 3D object in an image will change with the direction from which it was photographed. Thus, in the past it was necessary to prepare a large number of reference images beforehand. This technology enables more robust identification of 3D objects by automatically estimating the relative direction of shooting with respect to reference images, even from an image that was captured in such a way that the user could not see its front surface. From the viewpoint of service business operators, the number of images to be prepared beforehand can be significantly reduced since images need only be registered from a few directions. In addition, this technology makes it possible to identify a number of objects highly accurately, even if they are in an untidy environment or are partially concealed, by viewing them from feature points that satisfy constraint conditions on 3D objects derived from projective geometry.

(2) Subject identification technology using cloud data application

The optimal reference images are those in which only the subject is captured accurately and there is no unnecessary background. Our cloud data application subject identification technology (**Fig. 4**) is intended to achieve this through the use of our unique subject region extraction technology [2]. This technology makes it possible to identify and extract just the region in which the subject exists from a number of images that show the subject to be identified (**Fig. 5**). Application of this technology will make it possible to create reference images that show only their

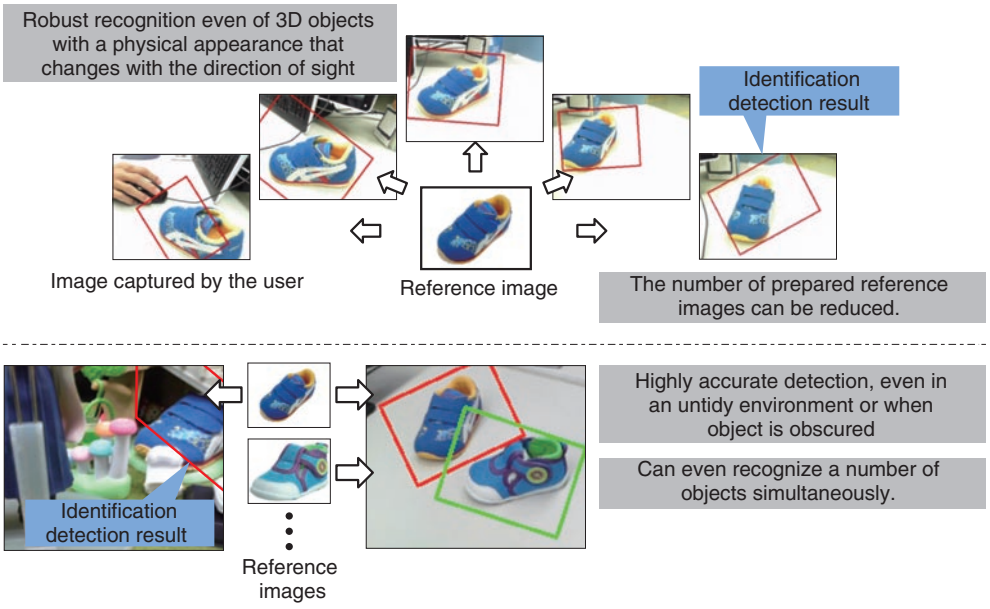

Fig. 3. Features of 3D subject identification technology.

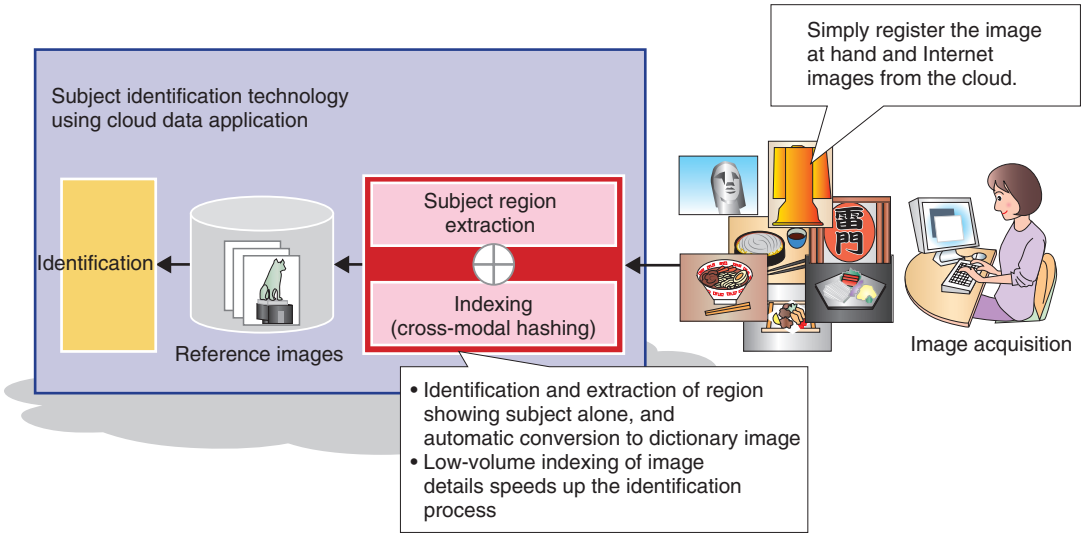

Fig. 4. Cloud data application type of subject identification.

subjects by simply registering the photograph at hand together with images acquired from the Internet. Not only does this lower the barrier to introducing and using the subject identification technology, but it is also expected to make it possible to increase the number of categories of subjects that can be identified by making it easy to prepare a huge number of reference images.

If it becomes possible to handle a large number of reference images, a rapid identification method that facilitates this will be essential. We are working on the R&D of unique indexing technology called *cross-modal hashing* [3]. This technology involves converting the content of each reference image into a very short code (hash) that is stored, making it possible to execute rapid identification processing by using the

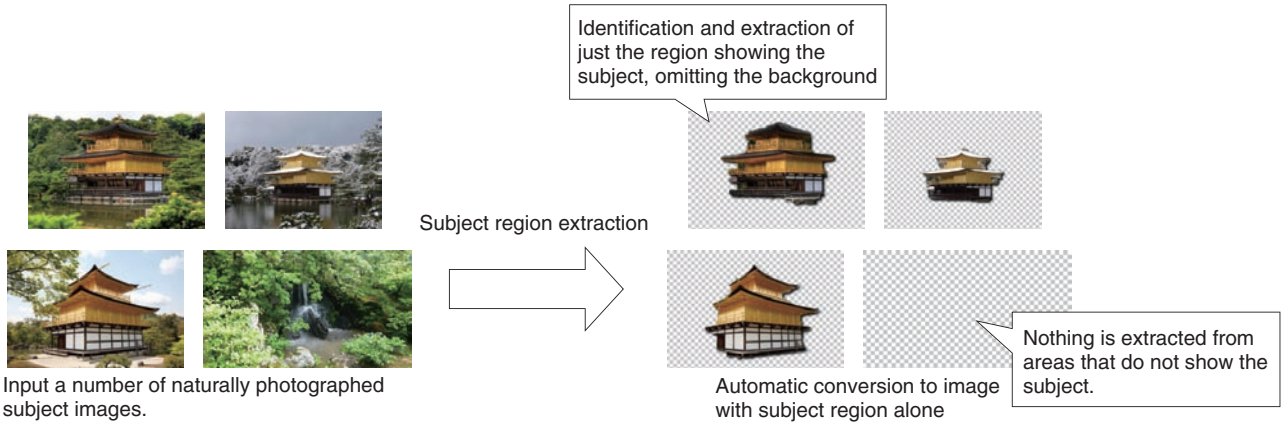

Fig. 5. Subject region extraction technology.

indexes, even with respect to a huge number of reference images. For example, a processing time of approximately 7 seconds was necessary in the past to identify subjects from a collection of 1 million reference images, but our technology has reduced that to less than half a second.

In the future, we will work on improving the technology to identify a large number of subjects rapidly and highly accurately from images, and we will continue to promote the implementation of user-friendly services that will enable smart assistance according to the user situation.

2.2 Natural language processing technology for understanding people and responding naturally

For an agent to behave naturally, it must have physical functions such as actions, but it is also important to have natural language processing technology in order to generate the content that the agent speaks. To that end, we introduce (1) user utterance intention comprehension technology, (2) automatic summary generation technology for generating natural descriptive sentences, and (3) interactive chat technology for putting together the entire interaction in a natural manner.

(1) Utterance intention comprehension technology

An overall image of the agent system is shown in **Fig. 6**. In this case, the description deals with the example of a scenario in which the agent plays the part of a sightseeing guide for a grandfather and grandchild.

The first thing that the agent should do is comprehend the intentions behind the contents of user utterances. Since a computer cannot comprehend human language, we are developing technology that roughly translates human language into a computer language (such as a database query language) [4]. For example, if the sentence "I wonder when it was built" and meta-information such as the current location can be translated into computer language as "s = Dendenji, p = year of construction, o = ?", it is possible to respond with "Dendenji was built in 1252."

(2) Automatic summary generation technology

However, simply returning single- phrase responses as described above does not make for a natural agent. We are therefore conducting research into improving the system's naturalness and intelligence by summarizing descriptive sentences as appropriate and adding them to the responses. Our approach is to take text information relating to a certain subject (Dendenji) that already exists on the web, convert it into colloquial form as a description, and provide information to be uttered by the agent. During this process, our automatic summary technology [5] plays a large role. When converting text on the web into a suitable form as a description, we use the automatic summary technology to generate a concise text passage with the redundant parts of the original text removed, and provide a natural description by converting it into colloquial form. This makes it possible to implement an inexpensive agent that provides natural descriptions. An example of the process when text on the web is converted into more colloquial-seeming speech is shown on the right side of Fig. 6. For example, the text passage: "According to the temple's history, in

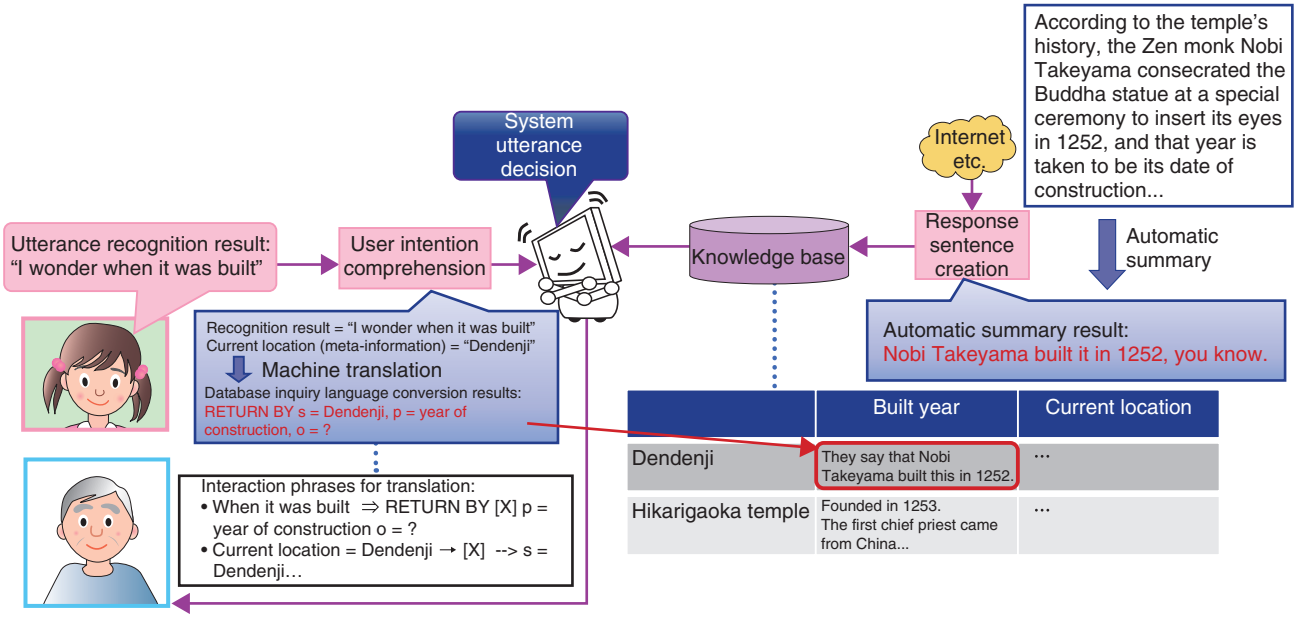

Fig. 6. Depiction of agent system.

1252 the Zen monk Nobi Takeyama* consecrated the Buddha statue at a special ceremony to insert its eyes, and that year is taken to be its date of construction..." can be used as an utterance by converting it into the brief, colloquial expression: "They say it was built by Nobi Takeyama in 1252."

(3) Interactive chat technology

It is possible to improve the quality of conversation by providing worthwhile information within the conversation, but to improve the naturalness of the conversation overall, it is necessary to improve the coverage of topics. Therefore, what we need is a chat function. A chat function is difficult to write in algorithmic form, though, and up until now, this kind of function has been implemented using hand-made rules. However, methods that rely on rules are expensive, and their coverage of topics is low. That is why we have constructed an interaction system that can chat automatically on a wide range of topics by turning text data on the Internet into interaction knowledge using language processing technology [6, 7] (**Fig. 7**). The system generates utterances from the predicate argument structure data (structure based on sentences formed of subjects, objects, etc.), based on the current topic and the utterance intentions of the system, and responds with a wide range of topics by selecting utterance sentences from an utterance database. It is also possible to make the phraseology suit the system characteristics by using the sentence-end expression conversion function.

Thus, natural language processing technology is essential for implementing an advanced interaction agent. In the future, we will continue with our R&D aimed at attaining a deeper understanding of language and on generating language for different objectives.

2.3 Speech synthesis technology capable of generating synthesized speech of a wide variety of speakers and speaking styles

Up until now, the main uses for speech synthesis technology were in information provision services such as those for verbally confirming safety information and those providing automated voice guidance systems in call centers. However, usage purposes have recently expanded to include applications such as voice interaction agents due to the recent spread of high-performance mobile terminals. Since the objective of information provision services is to convey information accurately by voice, it was sufficient to generate a synthesized speech with a single speaker. In contrast, with voice interaction agents, the requirement is now for synthesized speech with a wider

* This temple and the monk are fictional.

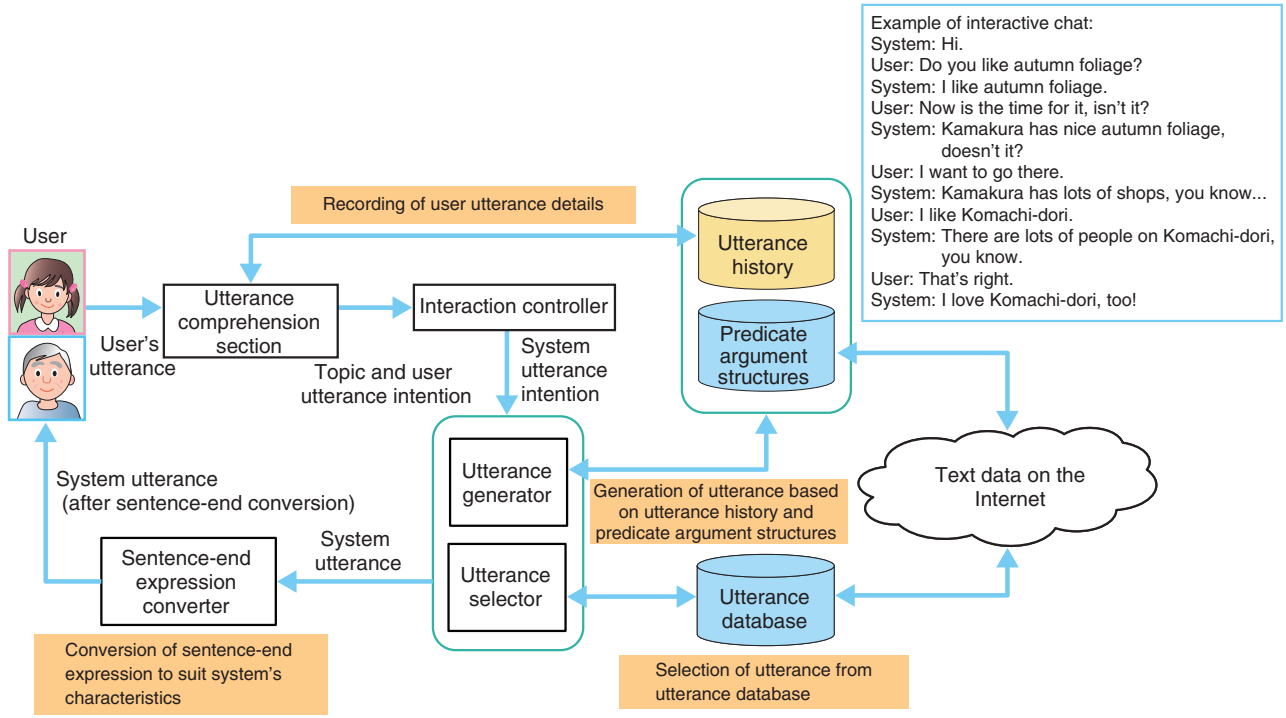

Fig. 7. Processing flow of interactive chat technology.

variety of speakers and speaking styles than in the past, for example, synthesized speech of various speakers that correspond to agents' characters, and synthesized speech with speaking styles that correspond to scenes that induce emotional expressions. We introduce here the novel speech synthesis technology that was developed by NTT Media Intelligence Laboratories for just such usage scenes, which enables the generation of synthesized speech for a wide variety of speakers and speaking styles.

An overview of the speech synthesis system is shown in **Fig. 8**. This technology consists of a training component that trains a model from an arbitrary speaker's speech and retains the characteristics of that speaker's speech, and a speech synthesis component that generates synthesized speech using that trained model.

In the training component, speech data uttered by a specified speaker are recorded. Then a model that has the speaker characteristics (voice quality and speaking style) of the target speaker is trained from the recorded speech of that speaker. The trained model consists of two parts: a speaker model that holds information on the voice quality of that speaker, and a style model that holds information on the speaking style of the speaker, for example, voice pitch and speaking rate.

In the speech synthesis component, a segment of synthesized speech that has the voice quality and speaking style of the speaker is generated from the trained speaker model and style model of the speaker. This technology also makes it possible to impart speaking styles to synthesized speech from within the style model that was trained beforehand, for example, a butler-style speaking style or a dramatic-reading-style speaking style. This means that when the speech synthesis is processed, it is possible to generate synthesized speech that has been given a specific speaking style while still having the voice quality of that speaker (such as speech with a dramatic-reading-style of speaking but with the voice quality of Ms. A).

One feature of this technology is that a shorter time for recording the speech of a speaker is necessary when implementing the speech synthesis of an arbitrary speaker. In order to generate synthesized speech having sufficient quality (reproducibility of speaker characteristics and naturalness of the synthesized speech) using conventional speech synthesis technology, a huge amount of speech data uttered by the

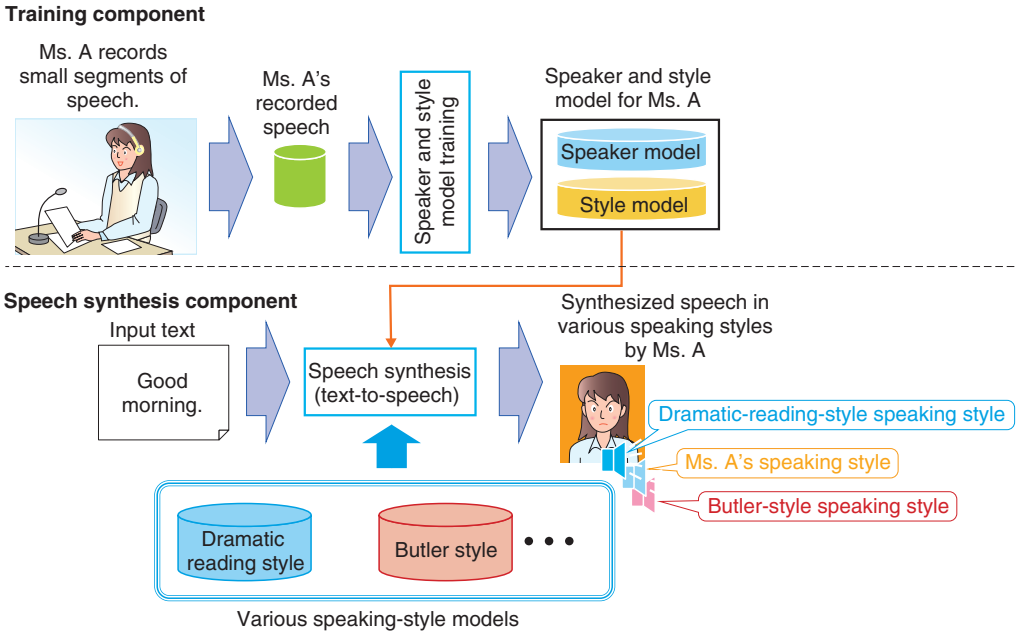

Fig. 8. Overview of speech synthesis system.

speaker are required. Since an extended speech recording taking between several hours to a dozen hours is necessary for collecting such speech data, it is difficult to create synthesized speech of various speakers because of the cost. This technology implements the simple creation of synthesized speech of various speakers corresponding to agents' characters, by greatly reducing the necessary speech recording time to between several dozen minutes and a couple of hours.

Our future task is to improve the basic performance such as the naturalness of the synthesized speech and the reproducibility of speaker characteristics. We will also investigate the speech synthesis required for the voice interaction interface, for example, the generation of synthesized speech with appropriate speaking styles for various usage scenes.

3. Future plans

Following our theme of hospitality in information search, we are focusing on issues that users encounter in their everyday lives and are working to involve agents in this process. We believe that further technical developments in the image recognition, language processing, and speech synthesis technologies introduced in this article will help achieve this.

In the future, we will pursue research on a new style of providing information naturally by using an approach that will help us understand potential preferences and cultural backgrounds, and by the entry of agents into user circles, with the aim of implementing services that will develop new relationships between users and agents that have not been seen before.

References

[1] J. Shimamura, T. Yoshida, and Y. Taniguchi, "Geometric Verification Method to Handle 3D Viewpoint Changes," MIRU2014 (the 17th Meeting on Image Recognition and Understanding), OS3-4, Okayama, Japan, Jul. 2014.
[2] S. Tarashima, G. Irie, H. Arai, and Y. Taniguchi, "Fast Web Image Object Cosegmentation with Region Matching," Proc. of Media Computing Conference 2014 (the 42nd Annual Conference of the Institute of Image Electronics Engineers of Japan), R4-2, Tokyo, Japan, Jun. 2014.
[3] G. Irie, H. Arai, and Y. Taniguchi, "Hashing with Locally Linear Projections," IEICE Transactions on Information and Systems (Japanese Edition), Vol. J97-D, No. 12, pp. 1785–1796, 2014.
[4] R. Higashinaka, K. Sadamitsu, W. Uchida, and T. Yoshimura, "Question Answering Technology in Shabette-Concier," NTT Technical Journal, Vol. 25, No. 2, pp. 56–59, 2013 (in Japanese).
[5] H. Nishikawa, K. Arita, K. Takana, T. Hirao, T. Makino, and Y. Matsuo, "Learning to Generate Coherent Summary with Discriminative Hidden Semi-Markov Model," Proc. of COLING 2014 (the 25th International Conference on Computational Linguistics), pp. 1648–1659, Dublin, Ireland, Aug. 2014.
[6] R. Higashinaka, "Towards an Open-domain Dialogue System," Proc. of the 70th SIG-SLUD (Special Interest Group on Spoken Language Understanding and Dialogue Processing) of the Japanese Society for Artificial Intelligence, Vol. 70, pp. 65–70, 2014.
[7] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi,

H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, “Towards an Open-domain Conversational System Fully Based on Natural Language Processing,” Proc. of COLING 2014, pp. 928–939, Dublin, Ireland, Aug. 2014.

Kugatsu Sadamitsu
Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.
He received the B.E., M.E., and Ph.D. in engineering from Tsukuba University, Ibaraki, Japan, in 2004, 2006, and 2009, respectively. He joined NTT Cyber Space Laboratories (now, NTT Media Intelligence Laboratories) in 2009. His current research interests include natural language processing and machine learning. He is a member of the Information Processing Society of Japan (IPSJ) and the Association for Natural Language Processing (NLP).

Jun Shimamura
Senior Research Engineer, Visual Media Project, NTT Media Intelligence Laboratories.
He received the B.E. in engineering science from Osaka University in 1998 and the M.E. and Ph.D. from Nara Institute of Science and Technology in 2000 and 2006, respectively. He joined NTT Cyber Space Laboratories (now, NTT Media Intelligence Laboratories) in 2000. From 2006 to 2010, he worked at NTT Communications. He received the 2005 TAF (Telecommunication Advancement Foundation) TELECOM System Technology Award. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Institute of Image Information and Television Engineers (ITE).

Go Irie
Research Engineer, Visual Media Project, NTT Media Intelligence Laboratories.
He received the B.E. and M.E. in system design engineering from Keio University, Kanagawa, in 2004 and 2006, respectively, and the Ph.D. in information science and technology from the University of Tokyo in 2011. He joined NTT Cyber Solutions Laboratories in 2006. He is a member of IEICE and ITE.

Shuhei Tarashima
Researcher, Visual Media Project, NTT Media Intelligence Laboratories.
He received the B.E. and M.E. in engineering from the University of Tokyo in 2009 and 2011. He joined NTT Cyber Solutions Laboratories in 2011. His research interests include machine learning and pattern recognition in computer vision, for example, object classification, detection, and segmentation.

Taiga Yoshida
Researcher , Visual Media Project, NTT Media Intelligence Laboratories.
He received the B.E. and M.E. in informatics from Kyoto University in 2007 and 2009, respectively. He joined NTT Cyber Solutions Laboratories in 2009 and studied video recommendation functions based on metadata and audio-visual features. He is currently studying object recognition. He is a member of IEICE.

Ryuichiro Higashinaka
Senior Research Scientist, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.
He received the B.A. in environmental information, the Master of Media and Governance, and the Ph.D. from Keio University, Kanagawa, in 1999, 2001, and 2008, respectively. He joined NTT in 2001. His research interests include building question-answering systems and spoken dialogue systems. From November 2004 to March 2006, he was a visiting researcher at the University of Sheffield in the UK. He received the Maejima Hisoka Award from the Tsushinbunka Association in 2014. He is a member of IEICE, the Japanese Society for Artificial Intelligence (JSAI), IPSJ, and NLP.

Hitoshi Nishikawa
Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.
He received the B.A. in policy management and the Master of Media and Governance from Keio University, Kanagawa, in 2006 and 2008, respectively, and the Ph.D. in engineering from Nara Institute of Science and Technology in 2013. He joined NTT in 2008. His research interests lie in the area of natural language processing, especially the study of automatic summarization. He received the Best Paper Award (first place) and Annual Meeting Excellent Paper Award from NLP in 2013 and 2014, respectively. He is a member of the Association for Computational Linguistics (ACL), NLP, JSAI and IPSJ.

Noboru Miyazaki
Senior Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.
He received the B.A. and M.E. from Tokyo Institute of Technology in 1995 and 1997, respectively. He joined the NTT Basic Research Laboratories in 1997. From 2004 to 2007, he was with NTT Cyber Space Laboratories (now, NTT Media Intelligence Laboratories). From 2008 to 2012, he was with NTT-IT Corporation. His current research interests include speech synthesis and spoken dialogue systems. He is a member of IEICE, the Acoustical Society of Japan (ASJ), and JSAI.

Yusuke Ijima

Researcher, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E. in electric and electronics engineering from the National Institution for Academic Degrees and University Evaluation upon graduation from Yatsushiro National College of Technology, Kumamoto, in 2007, and the M.E. in information processing from Tokyo Institute of Technology in 2009. He joined NTT Cyber Space Laboratories (now, NTT Media Intelligence Laboratories) in 2009, where he engaged in R&D of speech synthesis. He is a member of ASJ and IEICE.

Yukihiro Nakamura

Senior Research Engineer, 2020 Epoch-making Project, NTT Service Evolution Laboratories.

He received the B.E. and M.E. in engineering from the University of Tsukuba, Ibaraki, in 1992 and 1994, respectively, and the Ph.D. in functional control systems from Shibaura Institute of Technology, Tokyo, in 2014. He joined NTT in 1994. From 2001 to 2005, he was with NTT Communications. From 2011 to 2014, he was with NTT Advanced Technology Corporation. In 2014, he joined NTT Service Evolution Laboratories. His research interests include network robot platforms and human-robot interaction. He is a member of the Robotics Society of Japan (RSJ), the Japan Society of Mechanical Engineers (JSME), and the Society of Instrument and Control Engineers (SICE).