Speech Recognition Technologies for Creating Knowledge Resources from Communications

Yuichi Nakazawa, Takeshi Mori, Yoshikazu Yamaguchi, Yusuke Shinohara, and Noboru Miyazaki

Abstract

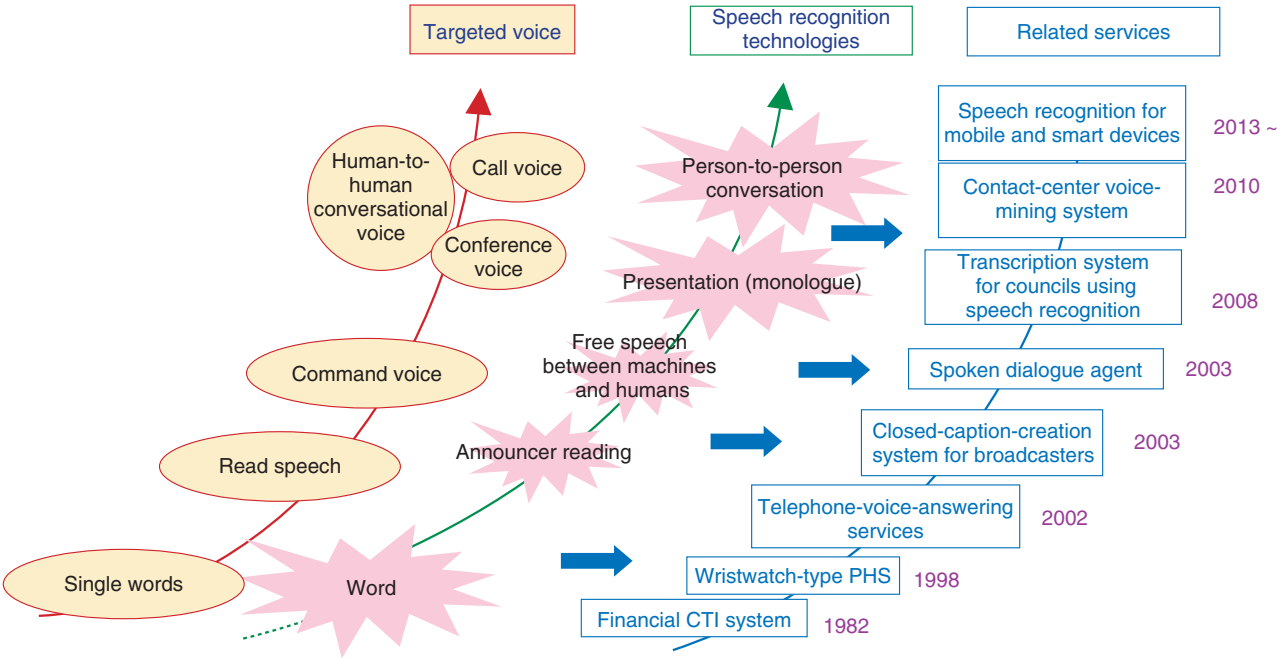
Speech recognition technologies have been used in an increasing number of situations to support or substitute work traditionally done by people such as call analysis in contact centers and support for creating minutes of council meetings. We are promoting research and development of speech recognition technologies because we believe that speech significantly supports human activities, especially in business. In this article, we describe the historical development of speech recognition technologies cultivated by NTT laboratories and the future contributions and roles of such technologies in human and corporate activities. It also introduces the use of non-linguistic information read from voice, such as emotion, gender, and age, which has been attracting attention.

Keywords: speech recognition, digital transformation, human-to-human communications

1. Development of speech recognition technologies

"Hey, Siri," "Ok, Google." These are the first words you speak to your voice assistant, and no doubt most people have used them. Speech recognition technologies have spread rapidly with the advent of voice assistants that control devices and provide information to people when speaking into a smartphone or artificial intelligence (AI) speaker. Various speech recognition technologies that enable such dialogue between people and computers have been put into practical use such as in automatic voice response devices (interactive voice response: IVR) in the 1980s and car navigation in the 1990s. However, with the recent introduction of deep-learning technologies, speech recognition accuracy has improved significantly, and the use of voice assistants and speech translation combined with machine translation due to increasing demand for it in the globalized world, has begun.

Nevertheless, speech is an important means of communication between people. Although the above-mentioned voice assistants work with relatively short utterances, commercialization of speech recognition for long utterances (long sentences, conversations) has also been studied. Since 2000, speech recognition technologies have been used in situations in which manuscripts were at hand and targeted speech was comparatively clear, such as captioning for news programs and support for council-minutes creation. However, the application of speech recognition technology for speech occurring in natural communications between people, such as analysis of operator-customer conversations at contact centers and real-time conversational support, has recently been advancing. Thus, speech recognition technologies have been developed with improved recognition accuracy and a wider variety of targeted speech (**Fig. 1**).



CTI: computer telephony integration
PHS: personal handy-phone system

Fig. 1. NTT speech recognition technology initiatives.

2. Role of speech recognition technologies in supporting digital transformation in business

We believe that by further expanding the range of speech targeted with speech recognition technologies, such technologies will be able to play a role in driving the transformation of corporate activities. Using information technology (IT), including AI technologies, to transform business processes is known as digital transformation (DX). The significance of DX is gaining attention regardless of industry or business type. However, advancing DX requires effort such as streamlining and automating business processes using IT and creating new value through seamless collaboration between business and IT. Speech recognition technologies have demonstrated their effectiveness in streamlining and automating business processes.

We use a large amount of speech when communicating with others. Although communications tools using text, such as social networking services, email, and chat, have advanced, real-time voice communications face-to-face or via telephone are probably more often chosen to confirm or convey complex content or for decision making that requires the consensus of many people.

Compared to text communications, voice communications are advantageous because they are in real time and convey unspoken information expressed through the nuances of voice. However, if recordings or notes are not made, information communicated by voice is volatile, i.e., as soon as it has been uttered it disappears. Although an enormous amount of voice communications occurs on a daily basis in corporate activities, currently only a limited amount of voice communications is targeted for data analysis such as calls between operators and customers at contact centers. The data in most voice communications are not used.

However, a large amount of voice communications that occurs face-to-face or via telephone between sales representative and customers contain information that could be useful from a variety of perspectives such as marketing or compliance management. Information from meetings and communications among employees (e.g., short consultations), is also useful for improving business activities. Examples of such information are seeds of new business ideas, business improvement tips, or employees' mental health conditions. Preserving such information by

turning it into text using speech recognition technologies to create knowledge resources for various processes, thus leading to business improvements, has the potential to contribute to the rationalization and automation of business processes. In the next section, we introduce example applications of speech recognition technologies to streamline business processes.

3. Use cases of speech recognition technologies

Compared to current speech recognition technology, research and development (R&D) in speech recognition technologies is now addressing broken utterances, which adds another layer of technical difficulty, enabling progress in business DX in an increasing number of situations. We introduce example use cases that are becoming more widespread due to improvements in speech recognition accuracy regarding broken utterances.

3.1 Conference speech recognition

Even though minutes are often taken in business meetings, many people who take minutes often feel that rather than quick notes, it requires much more work than predicted to actually record meetings. Attempting to faithfully take notes in a meeting means the minutes taker will not be able to properly participate in the meeting or its discussions or will worry whether the minutes will be accurately recorded sometime after the meeting. Also, recording all of a meeting and creating minutes while listening later takes more time than the meeting itself, so it might be a good to have an additional minutes taker participate in the meeting. Many people wish for such a device that can automatically and quickly create minutes.

Current speech recognition is not sufficiently accurate and is limited to searching speech tied to time information in the hope that important words can be recognized. However, in addition to supporting simple minutes creation, it is hoped that by enabling speech recognition for broken utterances, AI will become proficient in roles usually played by humans (facilitators), such as automatic creation of minutes in conjunction with summarization technology, linkage with issue-tracking systems by automatic extraction of action items, and facilitation of meetings.

3.2 Remote work support

Where remote operations and responses are expected to grow such as in healthcare and education, inconveniences that occur due to interactions not being face-to-face will have to be eliminated. In healthcare, for example, although it is technically possible to operate remote devices with buttons and levers as well as recording conversations, unobtrusive AI support technologies will be required so that the doctor, who only obtains a lower-than-normal amount of information about the patient through a screen, can focus on treatment. Such technologies could be used to pass a thermometer in response to "let's take your temperature" or automatically control the lighting inside the patient's mouth in response to "please open your mouth." Linking with dialect-conversion technology also holds promise for smooth communications in remote healthcare in rural areas where strong dialects are spoken.

In one-to-many classes, which are fundamental in education, student levels of understanding are understood through calling on all students for their responses. In online classes, however, it is clear that cross-talk occurring in such situations makes it difficult to establish voice communications. Unobtrusive AI that will prevent device operations from interfering with students' concentration, such as comprehending students' speech with real-time speech-to-text recognition and executing raised hand commands for student responses of "yes," will likely become as indispensable.

3.3 Contact centers

In contact centers, speech recognition technologies are being used for operator speech. If speech recognition results of customer speech can also be obtained with sufficient accuracy, speech recognition technologies could contribute to making business support more efficient, reducing operator tasks, increasing the number of calls that can be handled, and improving customer satisfaction to meet future increasing demand on contact centers as various services go online.

4. Use of non-linguistic information

As well as lingual information (text), non-linguistic information (gender, age, etc.) and para-linguistic information (emotions, intent, attitude) are contained in information conveyed through voice communications. Hence, there is demand to use such information to advance speech services in actual business.

We developed RexSense™, a software engine that can extract non-linguistic and para-linguistic information in speech, by studying technologies for recognizing and using non-linguistic and para-linguistic information as well as taking initiatives to recognize

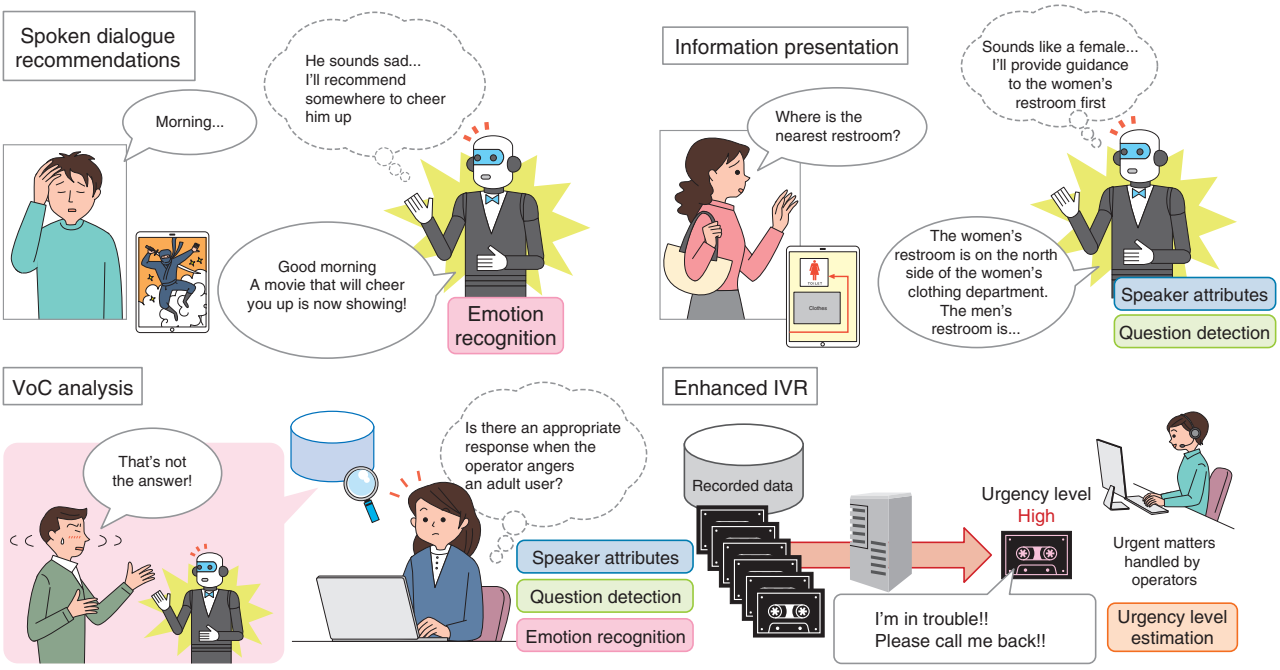


Fig. 2. RexSense™ application examples.

text information from speech with high accuracy. This software engine enables recognition and estimation of (1) speaker attributes (adult male, adult female, child), (2) emotions (joy, anger, sadness, calmness), (3) questions and non-questions, and (4) urgency, with high accuracy from voice data. We also developed the RexSense system to enable web application programming interface (Web API) services integrating this software engine with speech recognition for uses such as contact center advancement. RexSense also makes it possible to provide advanced services with robots by giving appropriate responses and recommendations according to human emotions or to provide advanced digital signage that presents more appropriate content (guidance, advertisements, etc.) based on non-linguistic information such as speaker attributes determined from voice. We expect the implementation of advanced voice of customer (VoC) analysis in contact centers, more sophisticated automated response services in IVR, and provision of more advanced audio conferencing solutions using non-linguistic and para-linguistic information (**Fig. 2**).

We also developed a customer-satisfaction-estimation technology for analyzing customer-voice characteristics and various conversation characteristics from the voices of operators in contact centers and customers to extract customer satisfaction (satisfaction and dissatisfaction) and introduced it to the contact center AI solution ForeSight Voice Mining™; commencing services in April 2019. We are developing a response-likability estimation technology for evaluating the likability of operator responses and considering putting it into service. These technologies hold promise for applications, such as call analysis (searching for good operator-response examples, analysis of customer satisfaction, etc.), operator support, operator and contact center evaluation, and operator education.

5. Future outlook

By expanding application areas from business scenarios and targeting all types of voice communications, the speech recognition technologies introduced in this article are crucial for achieving human Digital Twin Computing (DTC) [1], one element of the Innovative Optical and Wireless Network (IOWN) promoted by the NTT Group. In the cyber-physical interaction layer in the DTC architecture (**Fig. 3**), it will be necessary to collect the data required to generate digital twins by sensing real-space things (objects) and humans, and speech recognition technologies will play an important role in sensing human

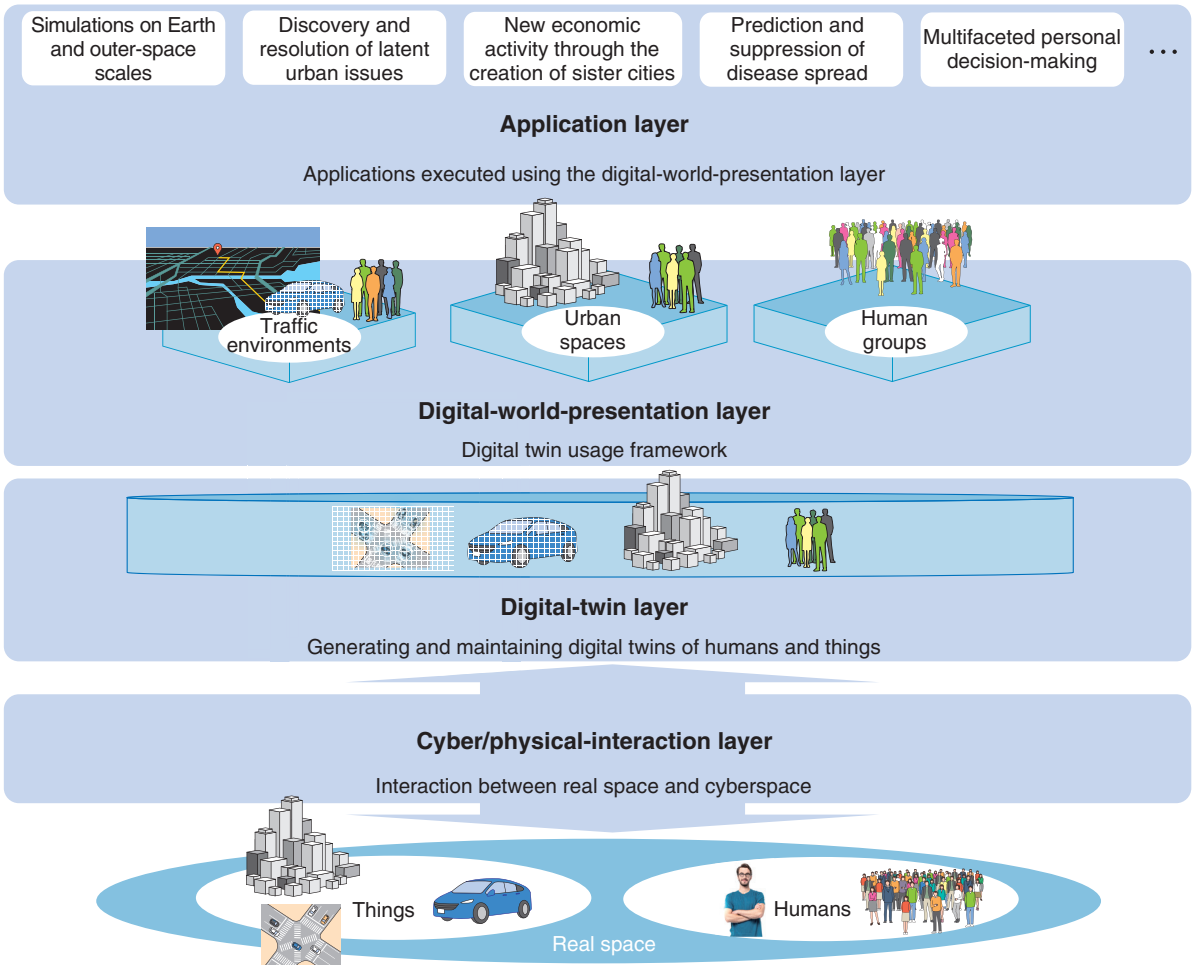


Fig. 3. Digital Twin Computing architecture.

thoughts.

Societies will become more convenient, richer, and safer as human DTC and DX of corporate activities progress. We will contribute to the creation of such a society through R&D of speech recognition technologies for human-to-human communications.

Reference

[1] I. Toshima, S. Kobashikawa, H. Noto, T. Kurahashi, K. Hirota, and S. Ozawa, "Challenges Facing Human Digital Twin Computing and Its Future Prospects," NTT Technical Review, Vol. 18, No. 9, pp.19–24, Sept. 2020.
https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202009fa2.html

Yuichi Nakazawa
Senior Research Engineer, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.

He received a B.E. and M.E. from Keio University, Kanagawa, in 2001 and 2003. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2003 and has been involved in R&D of automatic speech recognition technologies. He is a member of Acoustical Society of Japan (ASJ).

Takeshi Mori
Senior Research Engineer, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.

He received a B.E. and M.E. from Tokyo Institute of Technology in 1994 and 1996 and a D.E. from University of Tsukuba, Ibaraki, in 2007. Since joining NTT in 1996, he has been engaged in research on speech and audio processing algorithms. He is a member of the Institute of Electrical and Electronics Engineers (IEEE), ASJ, and the Institute of Electronics, Information and Communication Engineers (IEICE).

Yoshikazu Yamaguchi
Senior Research Engineer, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.

He received a B.E. and M.E. in electrical engineering from Osaka Prefecture University in 1993 and 1995. He joined NTT Human Interface Laboratories (now NTT Media Intelligence Laboratories) in 1995 and has been involved in R&D of automatic speech-recognition technologies.

Yusuke Shinohara
Senior Research Engineer, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.

He received a B.S. and M.S. in mechanoinformatics from the University of Tokyo in 2002 and 2004. From 2004 to 2017, he was a research scientist at Toshiba Corporate Research and Development Center, Kanagawa. Since joining NTT in 2017, he has been involved in R&D of speech recognition, in particular, acoustic modeling with various types of deep neural networks. His research interests include speech recognition, computer vision, and machine training. He received the IEICE Pattern Recognition and Media Understanding (PRMU) Young Researcher Award in 2004 and the ASJ Awaya Prize Young Researcher Award in 2013. He has served as a committee member (2013–2015, 2017-present) and secretary (2015–2017) for the Information Processing Society of Japan Special Interest Group (IPSJ SIG) on Spoken Language Processing. He is a member of IEEE, IEICE, ASJ, and IPSJ.

Noboru Miyazaki
Senior Research Engineer, Supervisor, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.

He received a B.E. and M.E. from Tokyo Institute of Technology in 1995 and 1997. Since joining NTT in 1997, he has been engaged in research on spoken-dialogue systems, speech synthesis, and speech-recognition technologies. He received the IEICE Inose Award in 2001. He is a member of ASJ, IEICE, and the Japanese Society for Artificial Intelligence (JSAI).