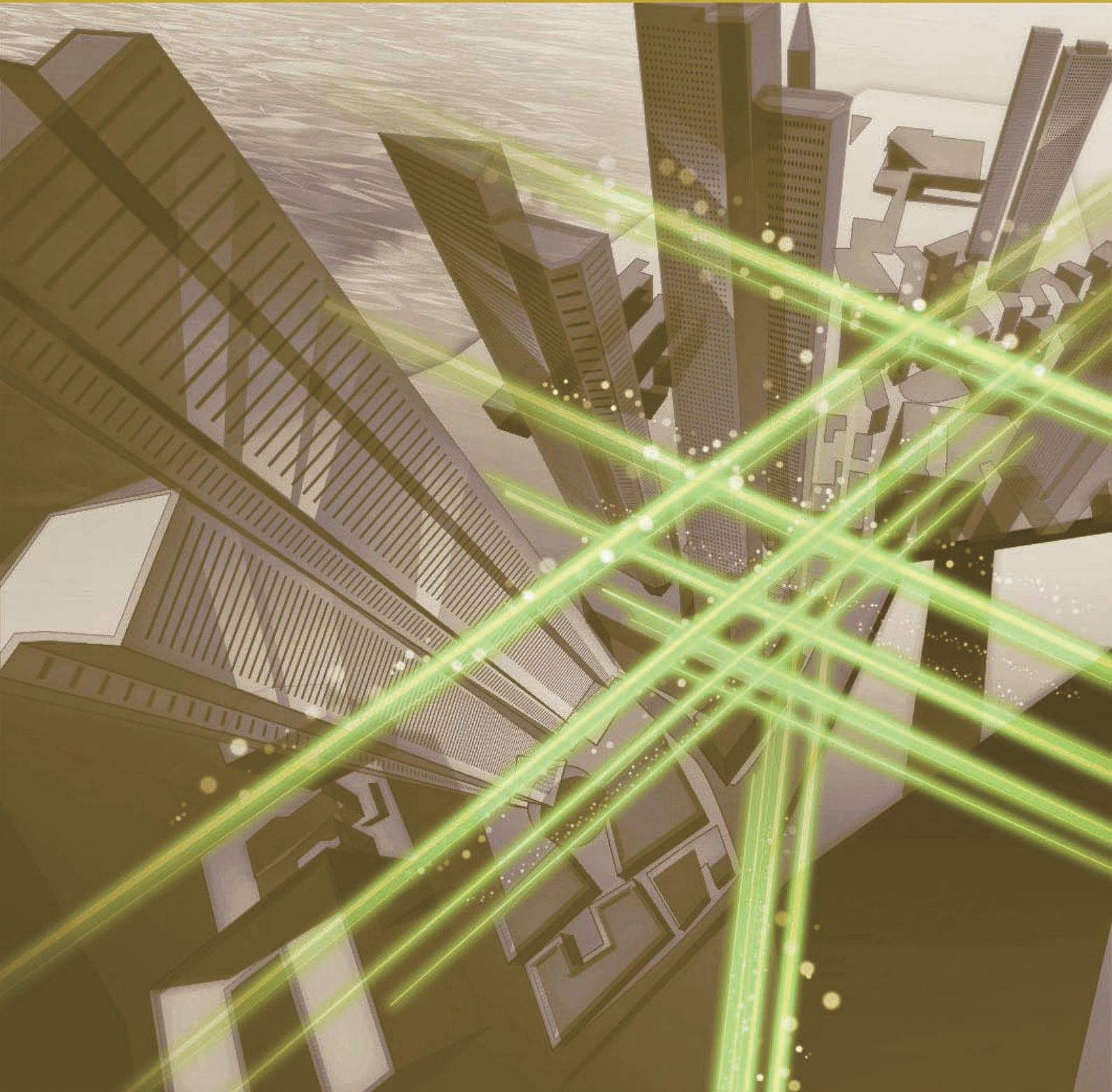


NTT Technical Review

2013

7



July 2013 Vol. 11 No. 7

NTT Technical Review

June 2013 Vol. 11 No. 7



Front-line Researchers

Masaaki Nagata
Senior Distinguished Researcher
NTT Communication Science Laboratories

Feature Articles: Intelligent Spoken Language Interface Technology for Various Services

Toward Intelligent Spoken Language Interface Technology

Speech Recognition Technology That Can Adapt to Changes in Service and Environment

Knowledge Extraction from Text for Intelligent Responses

Question Answering Technology for Pinpointing Answers to a Wide Range of Questions

Speech Synthesis Technology to Produce Diverse and Expressive Speech

Regular Articles

Multichannel Audio Transmission over IP Network by MPEG-4 ALS and Audio Rate Oriented Adaptive Bit-rate Video Codec

Distributed Array Antenna Technique for Satellite Communications

Global Standardization Activities

Development of ITU-T Action Plans for New Study Period at WTSA-12

NTT around the World

NTT Beijing Representative Office

External Awards

Committed to Easy-to-understand Explanations without Specialized Terminology



Masaaki Nagata
Senior Distinguished Researcher
NTT Communication Science
Laboratories

A translation machine that would allow anyone to communicate smoothly in real time with people in all sorts of countries and regions sounds like a fantasy, but the day that such a machine becomes a reality is not really that far away. Senior Distinguished Researcher Masaaki Nagata is a leading researcher of natural language analysis in Japan. We asked him to tell us about trends and issues in machine translation as well as his views on what being a researcher means.

World-leading statistical machine translation technology

—Dr. Nagata, please tell us about the research you are currently involved in.

Right now, I'm in charge of research conducted on statistical machine translation. This is a technology for achieving machine translation by using a huge amount of bilingual data compiled from previous translations to derive a statistical model corresponding to translation rules and a bilingual dictionary.

In general, the conventional machine translation process consists of analyzing a sentence in the source language, examining its structure, replacing words in that sentence with those in the target language, and reassembling the sentence in the target language. Specifically, this entails identifying parts of speech like nouns and adjectives and using the grammar of

the source language to determine the syntax of the sentence, that is, the subject and predicate, the main clause and subordinate clause, and so on. Then, once the syntax has been determined, a dictionary can be used to replace words in the source language with those in the target language. This would be a Japanese-English dictionary in the case of Japanese-to-English translation to replace Japanese words with English words. Finally, the word order in the source language, for example, subject-object-verb, must be changed to fit the grammar of the target language.

This method is implemented in line with grammatical and other types of rules and is therefore referred to as rule-based translation. In this scheme, a specialist creates translation rules based on his or her own knowledge and expertise. This kind of manual work, however, is extremely complicated, and it limits the accuracy that can be achieved. For this reason, “statistical machine translation” that automatically

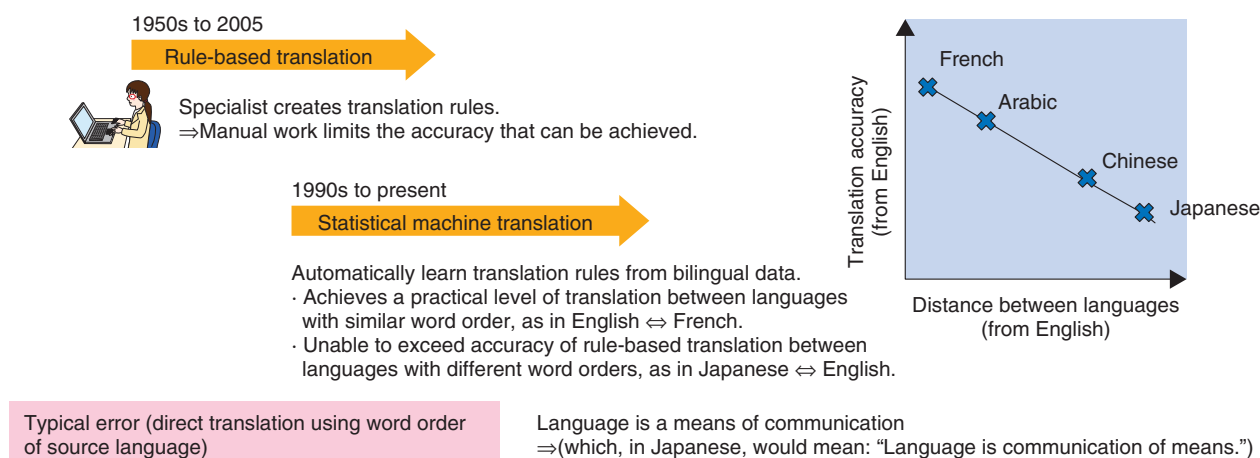


Fig. 1. Trends and issues in machine translation technology.

learns translation rules from a huge corpora of bilingual data has become the mainstream approach (Fig. 1).

This change from manually created rules to an automatically learned statistical model fits the recent pattern of technologically imitating human functions with computers, as in speech recognition and computer vision. It is difficult for rules created on the basis of human experience to be comprehensive, which is why this shift to statistical machine translation has taken place.

Another factor affecting translation accuracy is the distance between the languages in question. For example, grammar and word order in Korean are quite similar to those of Japanese, and as a result, meaning can be conveyed by simply replacing words using a bilingual dictionary. In contrast, English, which most Japanese people would probably need to have translated, can be ranked as the language furthest away from Japanese. French, meanwhile, is much closer to English, and high-accuracy machine translation between these two languages has come to be achieved relatively early. Translation is not very simple, however, between English and Japanese.

Recently, though, NTT proposed a technique that uses the "head-final" property of the Japanese language to translate English to Japanese after rearranging the English word order to that of Japanese. With this approach, we have succeeded in raising the translation accuracy.

—Can you explain this using a specific example?

Of course. Let me give you an example of translat-

ing an English sentence into Japanese using this technique.

In a phrase, which is a component of a sentence, the *head* is a word that determines the grammatical role of that phrase. In a prepositional phrase, for example, the preposition is the head word. To put it another way, the word in a phrase that is modified is the head. In this regard, the head-final property of Japanese means that the preceding word must modify the following word; that is, a modified word is almost always positioned toward the rear of the sentence with respect to its modifier. This property is rare among world languages.

In English, however, when a verb is present, the subject modifies it from the front while the object modifies it from the back. When a noun is present, moreover, an adjective modifies it from the front, and a preposition modifies it from the back. In the above sense, English has properties different from those of Japanese.

In accordance with the head-final property of the Japanese language, English words in the source document can be reordered so that modified words definitely come after their modifiers. In this way, the English word order is made to be the same as that in Japanese, and natural Japanese can then be achieved by simply performing a word-for-word translation (Fig. 2).

In summary, the preordering translation system that we have proposed obtains Japanese readings from English text according to the essentially same rules established for obtaining Japanese readings from classical Chinese text. This process consists of an

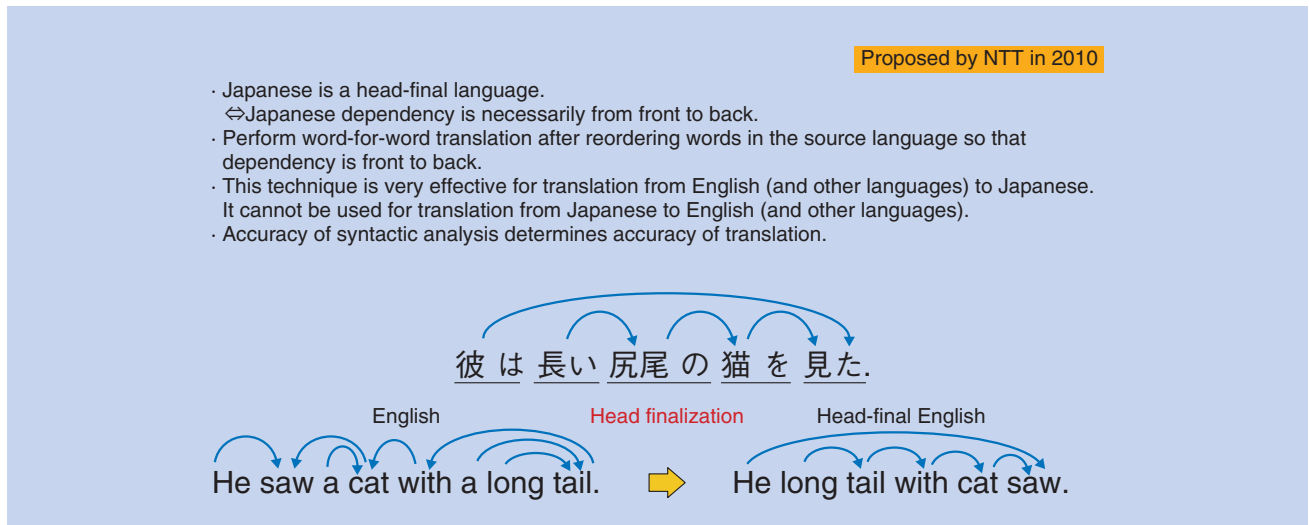
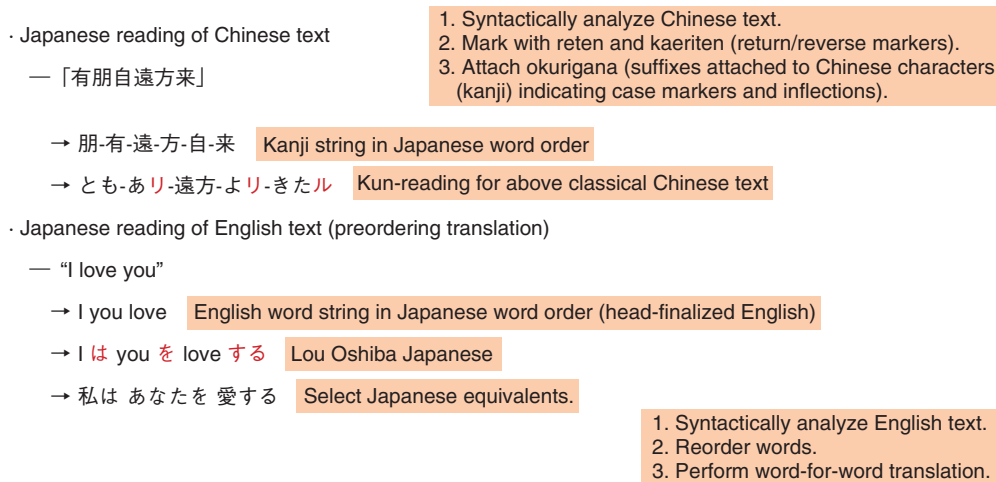


Fig. 2. Preordering based on Japanese head-final property.

Advanced statistical translation has reached a point corresponding to the 1000-year-old practice of obtaining Japanese readings from classical Chinese text.



有^レ朋^リ自^{ヨリ}遠^ニ方^ニ来^ル不^ニ亦^シ楽^シ乎[。]

Fig. 3. Japanese reading of Chinese text and preordering translation.

intermediate step in which English words are rearranged in the way that Lou Oshiba (a popular Japanese entertainer) speaks English using Japanese grammar and a final step in which that result is revised into correct Japanese (Fig. 3).

The problem here, however, is that, while we feel that translation to Japanese by this preordering technique has nearly reached a practical level, there are many unsolved problems in applying the technique to

Japanese-to-English translation.

At present, while we are working to improve the accuracy of English-to-Japanese automatic translation toward an actual product level, we are researching new systems for Japanese-to-English translation. One direction that future research of statistical translation and language analysis will take is establishing a technique for identifying the subject from a sentence that omits it, which is typical of the Japanese

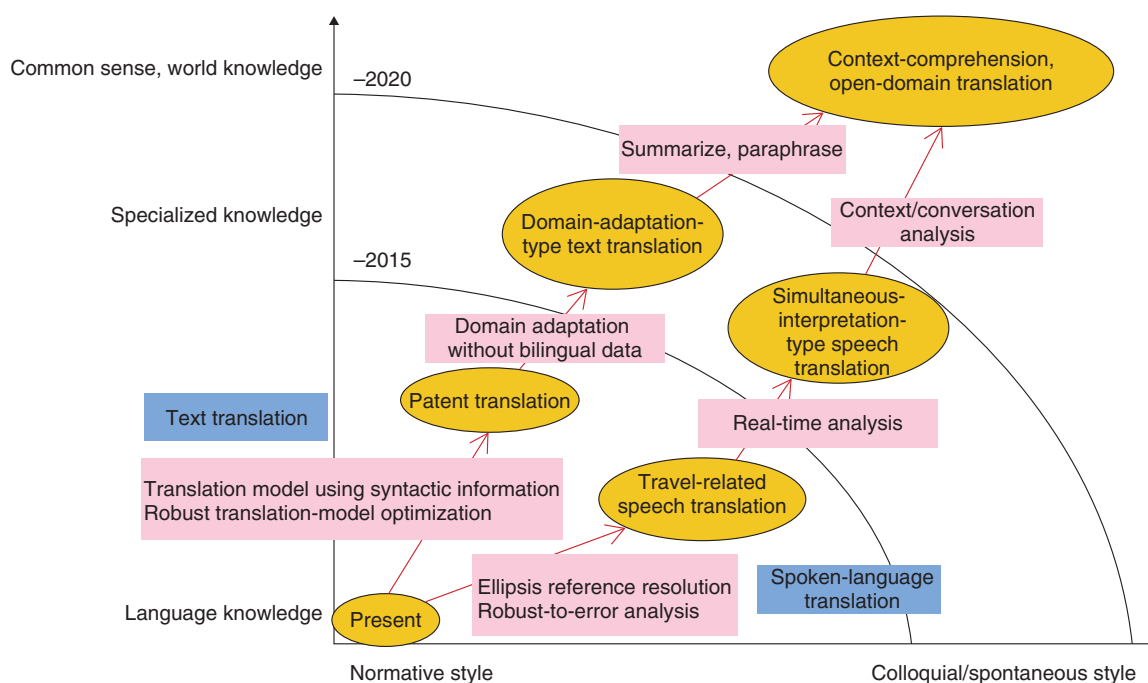


Fig. 4. Future research of statistical translation and language analysis.

language. This is the portion in **Fig. 4** indicated by “ellipsis reference resolution.” A function for translating “informal sentences” in the spoken language is difficult, but I think that translation of formulaic conversation as used, for example, in traveling, can be provided at a relatively early date. Beyond that, we envision a system that will enable smooth conversation with people who speak another language through real-time, that is, simultaneous, interpretation-type translation equipment. However, in the case of technical documents such as patent applications, I think we can reach a practical level in the near future.

Can a self-proclaimed language geek build a translation device by studying computer technology?

—How did you come to be involved in language analysis?

Actually, I was a “language geek.” Even today, whenever NHK begins a new language course, I cannot help but give that language a try. I wouldn’t say that speaking English is my forte, but my listening comprehension is good. I am capable of basic communication in Chinese, and during a time when I was studying Korean intensively, I was able to enjoy

popular Korean TV dramas without resorting to subtitles.

In my younger days, I enjoyed traveling overseas and visited more than 20 countries. French came in handy when I was lost in the old part of a Moroccan city and needed to ask directions, while German proved useful when I arrived in Prague by train late at night and had to make a hotel reservation by phone from the station. It is fun having unexpected experiences in strange countries that run counter to what one is used to in everyday life.

At university, however, I majored in information science. At that time, about 30 years ago, the objective was to develop people adept at creating both hardware (electronic circuits) and software, so what I studied is probably quite different from what today’s computer science students—who have probably never seen a soldering iron—are studying.

While in school, I was mainly involved in speech-related research. I worked in a laboratory that dealt not so much with language as with pattern recognition with the aim of implementing functions corresponding to human eyes and ears on the computer. Then, upon entering NTT, I was assigned to a group researching language processing, so I guess that would be the real beginning of my current research.

A demonstration style born out of a desire to have people appreciate research results

Why is research on language processing necessary? Today, language processing is being widely used in search sites, translation sites, and elsewhere on the Internet, so the need for research in this field should be easy to understand. However, at the time that I began my research, NTT had been a private company for only two or three years. It was a time when the business of NTT was centered around the telephone, that is, voice communications, and image communications such as with fax machines. Inside the company, the section that dealt with word and language processing focused only on telegrams and directory assistance. Language processing is a forward-looking type of research, so it was hardly mainstream at that time.

So, as to why I continued to research language processing, it's probably because I enjoyed it above and beyond the fact that this research was necessary. In the second half of the 1980s, soon after entering NTT, I was lent out to Advanced Telecommunications Research Institute International (ATR). I had wanted to do some innovative research, and I spent four years researching the creation of what is today called an "automatic interpretation telephone."

In that research, we constructed a system that combined speech recognition, machine translation, and speech synthesis as the world's first automatic interpretation telephone in a joint experiment with German and American universities. At that time, a demonstration of this translation system that we had constructed with enormous effort was receiving worldwide attention.

I learned two things from that experience.

To begin with, the research that I was actually involved in concerned the creation and input of rules by a specialist in so-called rule-based translation. I took up this research with much enthusiasm for the four years that I was at ATR. Nevertheless, we were not able to construct a very good system, and I became aware that it would be necessary to research the automatic learning of rules from language data.

Second, by presenting research results in the form of a demonstration, I found that I was really able to feel how people were responding. I found this to be very interesting. For a researcher, it is not enough to simply write a paper—he or she should lose no time in trying to move people with the results achieved. My former supervisor at university would always say, "It's not that 'examples are necessary, it's that

'examples are everything.'" Giving that demonstration reaffirmed those words in my mind and established my demonstration style for good.

Results must be conveyed to non-specialists in an easy-to-understand manner

—What is a typical day for you? Can you tell us something about your research style?

I am also a group leader, so I must divide my time between correspondence, meetings, and other responsibilities, which does not give me sufficient time to spend on research.

Nowadays, when listening to and discussing reports from researchers in our group, I often make decisions on our direction of research. And when reading papers and research reports in general, I like to consider what might be the next big thing in research. Furthermore, when writing a paper, I don't do anything special to be inspired, but I always strive to write in an easy-to-understand manner using good examples. Expressing principles as straightforwardly as possible is extremely important. As for criteria in assessing whether what I write is easy to understand, I am not aware of anything in particular other than asking myself whether people outside the specialized fields of the NTT laboratories will be able to understand my paper.

The NTT laboratories are involved in a variety of research themes, which means there is a wide range of specialized fields with each having an appropriate number of researchers. Specialized terms in one field are not necessarily understood in another field. If the significance of one's research is not understood, the significance of continuing one's research will not be recognized, and one's research results may never reach the implementation stage.

The desire to convey to the world what one is researching is common among researchers. Today, blogs and other Internet tools can be used for this purpose, but in my formative years as a researcher, this desire was satisfied by writing a textbook.

By the way, I have recently been involved in preparing presentations as a team member, and I participate in thinking about how best to convey our research results in an easy-to-understand manner to other people.

Additionally, I think we have a group of intellectually curious, energetic researchers here. Since I am older both in outlook and age, I would like to create an environment conducive to research for a young

generation of researchers. I would like to help up-and-coming researchers to broaden their outlook and work with them to devise methods of expression that make it easy for others to understand the nature of our research.

Having a passion to advance technology

—Dr. Nagata, it appears that you had one other turning point in your life as a researcher.

That's right. When I was around 40 years old, I developed a problem with my hip joint. I was informed by my physician that I had no choice but to rest and take it easy until I was 65. I really felt that my life as a researcher had come to an end.

However, thinking that there may be some other way to deal with my problem, I began to look through medical papers and discovered that a new type of surgery was available that could treat my condition. I then took it upon myself to find a doctor that could perform that surgery, and received the treatment. What occurred to me at that time was that medicine is also a world of research. There are various approaches, various schools of thought, and various opinions with respect to any one problem.

Thanks to advances in medical technology, I was able to return to work, and I thought then that I would like to contribute to society in my own research field. I became passionate about my work in a different way than before, and I resolved to work ardently toward the realization of automatic machine translation.

—Could you leave us with some advice for young researchers?

Stick with what you want to do and what you think is right without being influenced by what people

around you are saying.

Today, short-term results are needed in order to be recognized within the company; the idyllic atmosphere of the past is gone for good. In addition, I share the feeling that assessments from the outside world are becoming increasingly severe. Information now travels at lightning speed, and if you are wondering what other researchers are up to, you can find out in the blink of an eye. The environment today is completely different from that of the pre-Internet era. Today, having a self-centered attitude is unlikely, and it can be difficult to stick to one's beliefs. Nevertheless, I would say to researchers: "Maintain your sense of integrity in why you are pursuing certain research and where you are headed." I myself will support you as much as possible, so let's do our best!

Masaaki Nagata

Senior Research Scientist, Supervisor, Group Leader (Senior Distinguished Researcher), NTT Communication Science Laboratories.

He received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University in 1985, 1987, and 1999, respectively. He joined NTT in 1987. He was with ATR Interpreting Telephony Research Laboratories from 1989 to 1993. His research interests include natural language processing, especially morphological analysis, named entity recognition, parsing, and machine translation. He is a member of the Institute of Electronics, Information and Communication Engineers, the Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, the Association for Natural Language Processing, and the Association for Computational Linguistics.

Toward Intelligent Spoken Language Interface Technology

*Hirohito Inagaki, Takaaki Hasegawa,
Satoshi Takahashi, and Yoshihiro Matsuo*

Abstract

The spoken language interface has become much more prominent in recent years, as services for smartphones that take speech input for search functions and provide answers in the form of synthesized speech have grown in popularity. This article describes work being done at the NTT Media Intelligence Laboratories aimed towards implementing intelligent spoken language interface technology to support a variety of services.

1. Introduction

People are most accustomed to speaking their thoughts and intentions. With the rising popularity of applications and services for smartphones in recent years, various services that use speech as an interface for operating smartphones have been developed to supplement manual operation. Web search services in which speech can be used to input keywords to the Google search service or to smartphones that use the Google Android OS (operating system) are now available. Apple's iPhone is equipped with an interactive speech agent known as Siri. The interactive speech agent Shabette-Concier* provided by NTT DOCOMO for smartphones also has many users. Cell phones and smartphones are mobile devices that do not have keyboards, so it is not easy to input text quickly. Using a speech interface rather than a soft keyboard, touch panel, or other such manual input device also creates new value by making it possible to obtain various kinds of information from network databases while moving around outdoors or while engaged in indoor activities such as cooking that require the use of both hands. The naturalness and convenience of a speech interface is appealing because it is similar to the way people interact with each other. Nevertheless, easy access to networks and to various kinds of data and knowledge requires more than simple speech processing; it now also requires

new knowledge processing, natural language processing, and other back-end developments from Internet search technology. Furthermore, because a speech interface does not provide a visual medium, it is not suited to the immediate exchange of a large amount of information in the way that the screen of a personal computer is suitable for. Thus, accurately grasping what information the user is requesting, and selecting information on the basis of that understood request is even more important for a speech interface than when the information is output to an ordinary Internet browser on the screen of a personal computer.

The NTT laboratories are working to implement a personalized user interface and user experience (UI/UX) that are simpler and easier to use for the development of advanced services by linking information processing functions for various kinds of existing databases and Internet information in addition to developing front-end functions for highly accurate processing of speech input and output.

This article explains the most recent research on the intelligent spoken language interface technology that is needed to support the various increasingly advanced services of the future.

* "Shabette-Concier" is a registered trademark of the NTT DOCOMO Corporation.

2. Evolution of the intelligent spoken language interface

We begin by looking back on the history of dialogue systems, focusing particularly on their configuration, and then we consider the directions of intelligent spoken language interface development and the conditions it requires. Research on technology for the interaction of humans and computers has a long history. About half a century ago, the interactive system consisting of typing on a keyboard to communicate with a computer appeared. That was followed by research and development on speech interaction systems that understand spoken words. Around 2000 in the U.S., there was a boom in the development of practical voice portals and interactive voice response (IVR) systems, in which speech interaction processing was performed on a remote server. Telephones have now become smartphones, and speech interaction with computers anytime and anywhere has become possible through both wired and wireless connections. Furthermore, dialogue system servers can be accessed via the Internet, enabling the acquisition of knowledge from the immense collection of documents that are available on the Internet. That has made it possible to answer even broad questions for which it is difficult to prepare responses in advance. Such documents are continuously being added, and using them as a source of knowledge enables real-time handling of information in finer detail.

The continuing development of the spoken language interface must include further development of knowledge processing and natural language processing (the back-end processing) as well as development of speech recognition and synthesis processing that is suited to the user's context (the front-end processing).

We can view the spoken language interface as having two directions:

- (1) Human-agent spoken language interface (second-person interface)

This is an extension of current interactive systems in which an agent supports user thought and behavior by anticipating the user's intention on the basis of the user's present situation and behavioral history, gathering the precise information from the large amount of various kinds of data that are stored on the Internet, and composing an answer that is as complete as necessary.

- (2) Human-human spoken language interface (third-person interface)

This is an agent that can invigorate communication

between humans, whether face-to-face or over the Internet, by autonomously gathering information that is relevant to the topic of conversation and information from the communication history and profiles of the participants, and spontaneously offering useful information.

Both of these directions require personalization that involves strong awareness of the user at the front end. The NTT laboratories are researching spoken language processing as part of their efforts to realize such an intelligent spoken language interface.

3. Elemental technology for constructing an intelligent spoken language interface

The elemental technology needed to configure an intelligent spoken language interface is illustrated in **Fig. 1**. First, the user's speech is input to a speech recognizer and recognized on the basis of an acoustic model and a language model. The result is then converted to text. Next, the converted input is sent to a problem solver, which performs some processing and generates a response according to the result. This special issue specifically concerns the processing for generating an answer in response to a question and the processing for extracting the knowledge that is required to form an answer from a large collection of documents. These types of processing are positioned as applied natural language processing technology. In the final step of the intelligent spoken language interface, a speech synthesizer converts the textual output of the problem solver to speech using a speech database and prosody generation model.

The following sections describe the research on these technologies.

3.1 Speech recognition

The accuracy of speech recognition is extremely important in the intelligent spoken language interface because it strongly affects the overall performance of the system. We have therefore attempted to increase the accuracy of speech recognition in various ways. One approach is to achieve high recognition accuracy for unspecified speakers rather than particular individuals. The accuracy of speech recognition for particular users can be low, so it is necessary to improve recognition so that it is accurate for any user.

Another approach is to achieve good recognition accuracy when background noise is present. Use environments are diverse, and ambient noise is particularly problematic for mobile devices. At a train station platform or bus stop, for example, the sound

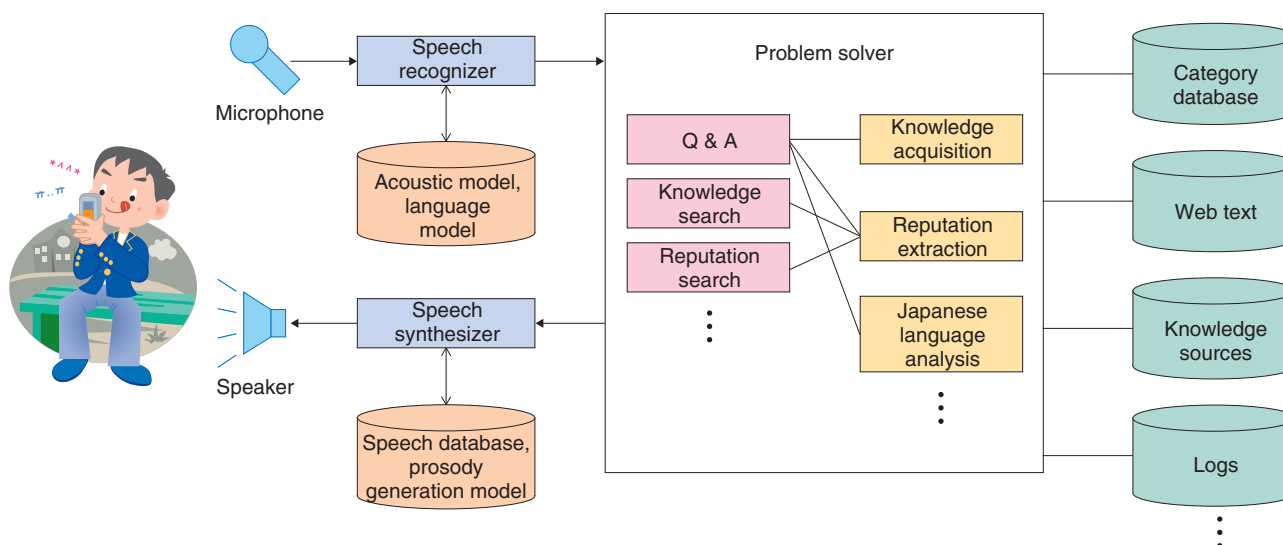


Fig. 1. Overview of an intelligent spoken language interface.

of passing trains or vehicles is present, and in restaurants, coffee shops, or other such locations where people gather, there is the noise of other people talking. There is thus a need for accurate recognition of user speech even under such noisy conditions.

A third approach to accurate speech recognition is the ability to recognize new words that enter the lexicon day by day. Generally, words that are not in dictionaries (referred to as out of vocabulary (OOV)) are a problem in speech recognition. The accuracy of a service can decrease greatly when OOV words cannot be recognized. It is therefore necessary to look for new words that should be recognized in past utterances of the user and other places.

New speech recognition technology for the three approaches described above is explained in detail in the article entitled “Speech Recognition Technology That Can Adapt to Changes in Service and Environment” [1] in these Feature Articles.

3.2 Knowledge extraction

The knowledge extraction performed by the problem solver extracts the information required to accurately answer a wide variety of user questions from a large set of Web pages. For example, if a user asks a question for which the answer is the name of a person or a mountain, the solver collects the names of persons and mountains from a relevant set of documents. Objects that have specific names to be referred to are called named entities (e.g., the mountain known as *Everest* is a named entity). Examples of named enti-

ties include places, organizations, time expressions, numerical expressions, and monetary amounts. In implementing an intelligent spoken language interface, *knowledge* is required in order to *understand* the meaning of user utterances and to generate responses. Named entities are an important element in constructing such knowledge. For example, named entities can be used to construct a particular individual’s information. For example, if it is possible to extract the name of a certain person, then attributes such as that person’s date of birth (time expression) and place of birth (place name) could be included in the response. If it were also possible to extract the relations among named entities, then correct answers to even more complex questions would be possible. Technology for extracting named entities and the relations among them from a large set of Web pages is described in the article entitled “Knowledge Extraction from Text for Intelligent Responses” [2].

3.3 Question answering

In the question answering process performed by the problem solver, the user’s question is first analyzed to determine what the user is asking and what type of question it is (who, what, where, or when) in order to determine what specific kind of named entity the question concerns. For example, is the user asking for a person’s name, the name of a mountain, or the name of a food, etc.? Then, various calculations are performed to select the most suitable candidate for the content of the question from among the named

entities extracted from Web pages in order to form the answer. The major difference between question answering and document retrieval is that the response to question answering is a pinpointed answer rather than a list of documents, so very high accuracy is expected in the answer. The question answering technology, which is the heart of the intelligent spoken language interface, is explained in the article entitled “Question Answering Technology for Pinpointing Answers to a Wide Range of Questions” [3].

3.4 Speech synthesis

Speech synthesis technology generates speech with a synthetic voice of good quality by reading text correctly and with the appropriate intonation and pauses. We are also researching various aspects of speech synthesis for implementing an intelligent spoken language interface. In one line of research, the user is allowed to freely select the voice of a person they would like to hear, such as a family member or friend. Being able to easily synthesize voices in that way could lead to the implementation of services that can be adapted to individuals. Another area of investigation involves the generation of clearly articulated speech that can be heard clearly even in the presence of background noise in crowded user surroundings. Speech synthesis technology that implements these features is explained in the article entitled “Speech Synthesis Technology to Produce Diverse and Expressive Speech” [4] in these Feature Articles.

4. Future work

We have presented an overview of speech recognition, speech synthesis, and natural language processing (knowledge extraction and question answering), all of which constitute the intelligent spoken language interface. Further progress requires audio signal processing for advanced sound acquisition and reproduction technology for mobile devices in addition to development of the technologies we have described. We will continue with the research and development of audio, speech, and language media technology that integrates these technologies with the objective of realizing a personalized UI/UX.

References

- [1] H. Masataki, T. Asami, S. Yamahata, and M. Fujimoto, “Speech Recognition Technology That Can Adapt to Changes in Service and Environment,” NTT Technical Review, Vol. 11, No. 7, 2013. <https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa2.html>
- [2] K. Sadamitsu, R. Higashinaka, T. Hirano, and T. Izumi, “Knowledge Extraction from Text for Intelligent Responses,” NTT Technical Review, Vol. 11, No. 7, 2013. <https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa3.html>
- [3] R. Higashinaka, K. Sadamitsu, K. Saito, and N. Kobayashi, “Question Answering Technology for Pinpointing Answers to a Wide Range of Questions,” NTT Technical Review, Vol. 11, No. 7, 2013. <https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa4.html>
- [4] H. Mizuno, H. Nakajima, Y. Ijima, H. Kamiyama, and H. Muto, “Speech Synthesis Technology to Produce Diverse and Expressive Speech,” NTT Technical Review, Vol. 11, No. 7, 2013. <https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa5.html>



Hirohito Inagaki

Vice President, Director, NTT Media Intelligence Laboratories.

He received the B.E. and M.E. degrees in electrical engineering from Keio University, Kanagawa, in 1984 and 1986, respectively. Since joining NTT Electrical Communication Laboratories in 1986, he has been engaged in R&D of natural language processing and its applications. From 1994 through 1997, he was on transfer to NTT Intelligent Technology Co., Ltd., where he was engaged in developing information security systems and multimedia systems. He moved to NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2006. His research interest is R&D of audio, video, and language interfaces and their applications. He is a member of the Institute of Electronics, Information and Communications Engineers (IEICE).



Takaaki Hasegawa

Senior Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E. and M.E. degrees from Keio University, Kanagawa, in 1992 and 1994, respectively, and the Dr.Eng degree from Tokyo Institute of Technology in 2010. Since joining NTT in 1994, he has been engaged in the research of natural language processing and intelligent information access. He was a visiting researcher at New York University from 2003 to 2004. He is a member of the Information Processing Society of Japan (IPSI), the Japanese Society for Artificial Intelligence, and the Association for Natural Language Processing (NLP).



Satoshi Takahashi

Executive Manager, Executive Research Engineer, Supervisor, Audio, Speech and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E., M.E., and Ph.D. degrees in information science from Waseda University, Tokyo, in 1987, 1989, and 2002, respectively. Since joining NTT in 1989, he has been engaged in speech recognition, spoken dialog systems, and pattern recognition. He is a member of the



Yoshihiro Matsuo

Group Leader, Senior Research Engineer, Supervisor, Audio, Speech and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.S. and M.S. degrees in physics from Osaka University in 1988 and 1990, respectively. He joined NTT Communications and Information Processing Laboratories in 1990. He moved to NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2001. His research interests include multimedia indexing, information extraction, and opinion analysis. He is a member of IPSJ and NLP.

Speech Recognition Technology That Can Adapt to Changes in Service and Environment

Hirokazu Masataki, Taichi Asami, Shoko Yamahata, and Masakiyo Fujimoto

Abstract

Speech recognition is a technology that recognizes spoken words. Speech recognition is a fundamental function for speech and language interfaces, and its quality strongly affects interface usability. Therefore, the recognition accuracy should always be high. However, recognition accuracy can drop significantly if there are changes in the service or the use environment. In this article, we introduce research being carried out to tackle this problem.

1. Introduction

1.1 User interface using speech recognition technology

Speech recognition allows machines to listen to our speech and convert it to text; the technology is thus roughly equivalent to the human ear. As speech recognition is a major component of speech and language interfaces, and its quality strongly influences the usability of the interfaces, the perennial demand is for high recognition accuracy. The NTT laboratories have been researching speech recognition technology for more than 40 years. Around the year 2000, speech recognition became practical in ideal environment owing to the many years of technological innovation and the advances in computer performance [1], [2].

The standard architecture of a speech recognition system is shown in **Fig. 1**. The technology is composed of feature extraction, an acoustic model, and a recognition dictionary. Existing speech recognition systems can recognize our voice only when uttered in a quiet environment using standard voice quality and common words. Consequently, recognition accuracy is significantly degraded in practical use.

1.2 Elements that can cause degradation

(1) Noisy environments

If you speak in a noisy environment such as a busy station or other crowded place, recognition accuracy will be poor because speech features are distorted by noise.

(2) Weak speakers

Even if the acoustic model is trained using the speech from a large number of speakers, recognition performance is poor if the speaker's speech features differ widely from those used to train the acoustic model.

(3) Out-of-vocabulary (OOV) failures

It is impossible to recognize words that are not in the dictionary with current speech recognition technology. This includes newly coined words and infrequently used words.

In this article, we introduce recent research advances that tackle these problems.

2. Voice activity detection and noise suppression

Ambient noise seriously degrades speech recognition accuracy. Thus, effective techniques for detecting voice activity and suppressing noise are critical for improving speech recognition accuracy in noisy environments. Voice activity detection accurately

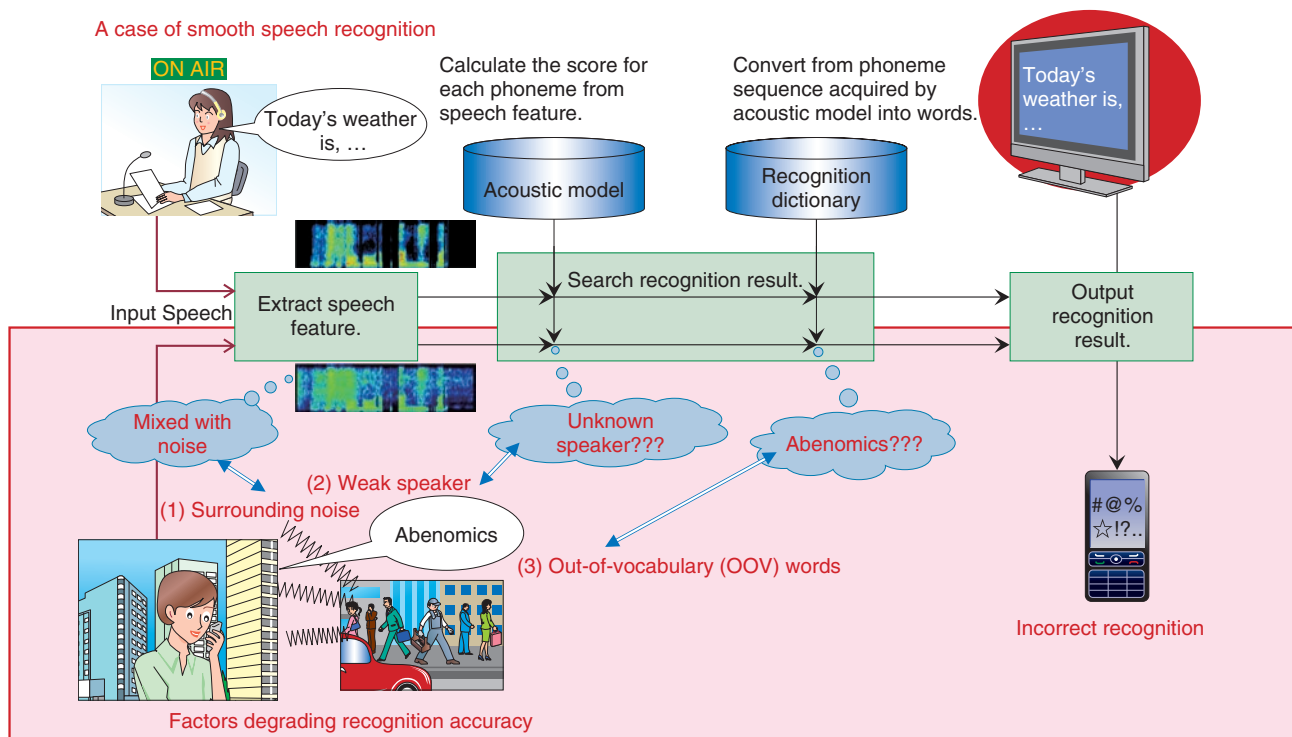


Fig. 1. Principle of speech recognition and problems.

detects periods of human utterances in noise-corrupted speech signals, and noise suppression clearly extracts signals of human utterances from the signals of the detected periods by suppressing the noise components. Typically, these techniques are individually developed and added as discrete front-end processes of a speech recognition system; the output of one is the input of the other in a simple chain. However, since these individual processes cannot share important information, it is difficult to achieve advanced front-end processing. A key problem is the accumulation of errors created by each technique.

To address this problem, we developed an integrated technique for voice activity detection and noise suppression called DIVIDE (Dynamic Integration of Voice Identification and DE-noising). DIVIDE reduces the number of processing errors in both processes and achieves front-end processing with advanced performance. DIVIDE employs statistical models of human speech signals and uses the models in both voice activity detection and noise suppression, as shown in Fig. 2. In DIVIDE, the activity probability of a human utterance is calculated from the statistical models in each short time slice.

When the activity probability exceeds a certain

threshold, the time slice is tagged as a period of a human utterance. The tagging allows the noise components to be suppressed. Namely, DIVIDE extracts discriminative information that represents the similarity of human speech by utilizing statistical models of human speech signals as a priori knowledge in front-end processing. With this information, voice activity detection and noise suppression are performed simultaneously in DIVIDE.

Speech recognition results in noisy environments are shown in Fig. 3. The graph reveals that DIVIDE considerably improves speech recognition accuracy.

3. Automatic adaptation to weak speakers

Acoustic models that map our speech to phonemes* are used in automatic speech recognition. The mapping between speech and phonemes is acquired from manual transcriptions of speech by using machine learning methods ((1) in Fig. 4). We preliminarily train the acoustic model by using the speech samples of many people because each person's speech is

* A phoneme is the minimum unit of a speech sound and roughly equivalent to the sound corresponding to a single letter of a Japanese word written in Roman letters.

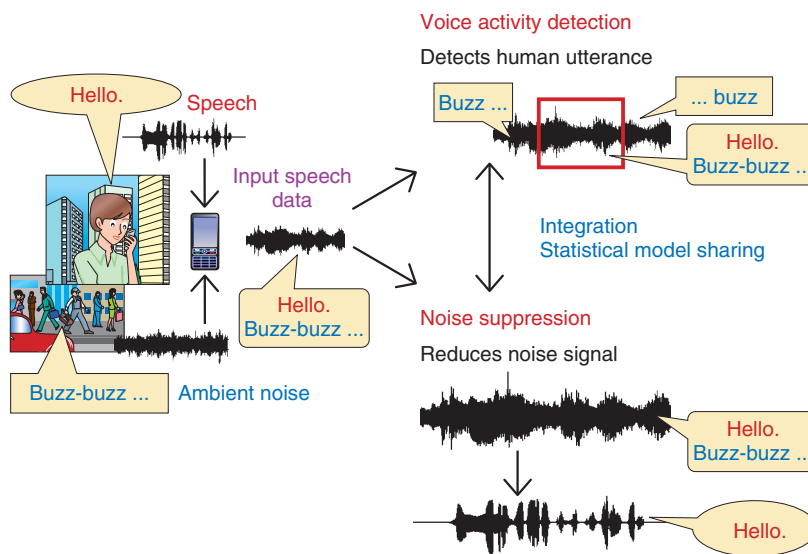


Fig. 2. Overview of DIVIDE.

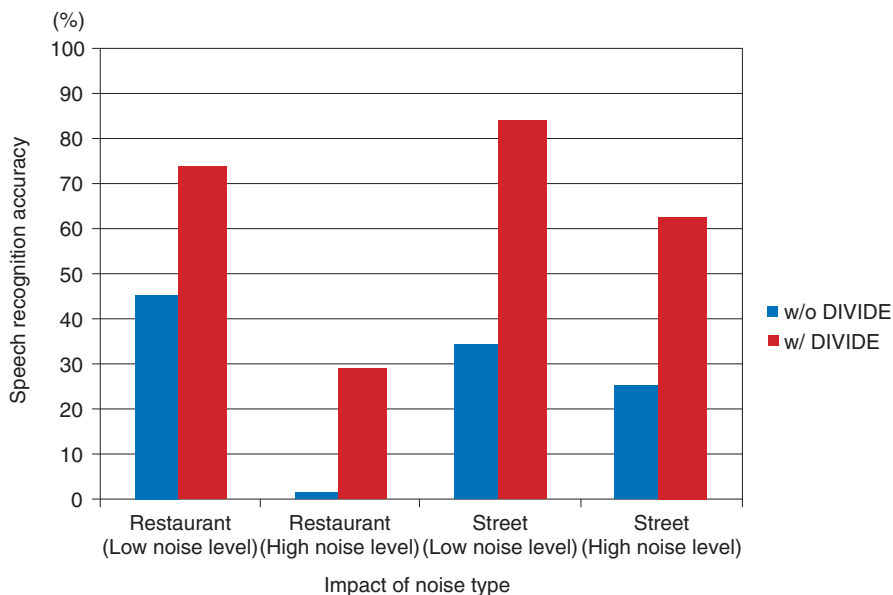


Fig. 3. Results of speech recognition using DIVIDE.

different. Even if the user’s speech has not been learned, the acoustic model trained by similar speech can map the user’s speech to the correct phoneme. Thus, the speech of many users can be accurately recognized.

However, even if learning involves many speakers, it is impossible to eliminate all blind spots. In actual use, there are always some users who are classified as

weak speakers in that the acoustic model cannot accurately recognize their speech.

To solve this problem, we have developed a method of automatic adaptation to weak speakers. This method allows in-service speech recognition engines to identify weak speakers, automatically learn their speech, and maintain high recognition accuracy (Fig. 4).

Our speech recognition engine has a function that

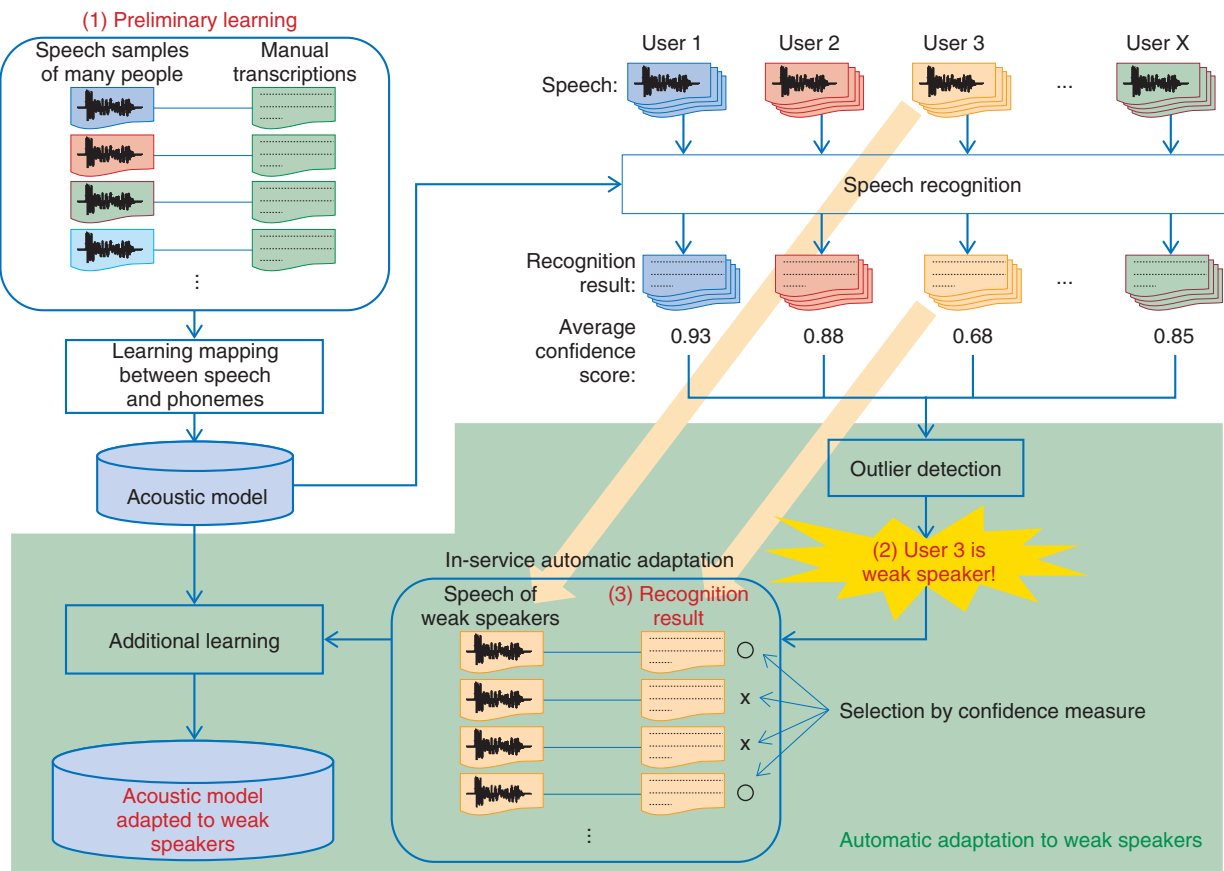


Fig. 4. Overview of automatic adaptation to weak speakers.

grades the correctness of its own recognition results [4]. This confidence score allows the engine to automatically find weak speakers. When the confidence score of a particular user is so low as to be an outlier, the user is taken to be a weak speaker ((2) in Fig. 4).

For the acoustic model to learn the mapping between the speech and phonemes of weak speakers, the direct solution is to obtain manual transcripts of their speech. However, creating manual transcripts is too expensive in terms of cost and time. Thus, our method uses the recognition results instead of manual transcripts to automatically train the acoustic model ((3) in Fig. 4). Unlike manual transcripts, the recognition results include erroneous parts that differ from the actual speech content; this makes learning ineffective. To address this problem, we turn to confidence scores again. Well-recognized results are selected by thresholding the confidence scores and then used for learning. We confirmed through experiments that our method improved the recognition accuracy of 80% of weak speakers to the same level

as regular speakers.

4. Automatic vocabulary adaptation

Speech recognition systems include a *recognition dictionary*, which is a list of words to be recognized. All recognition dictionaries are limited in terms of coverage because of cost concerns, so OOV word failures are unavoidable. The current solution is to update the dictionaries manually by adding new words such as the names of new products or the titles of new books.

Several methods have been proposed to avoid having to manually update dictionaries. These methods collect web documents related to user's utterances, extract OOV words from the relevant documents, and add the OOV words to the recognition dictionary (Fig. 5(a)). However, these methods register all OOV words in the relevant documents, even words that are not spoken in the target utterances (i.e., redundant word entries). Consequently, useful words may be

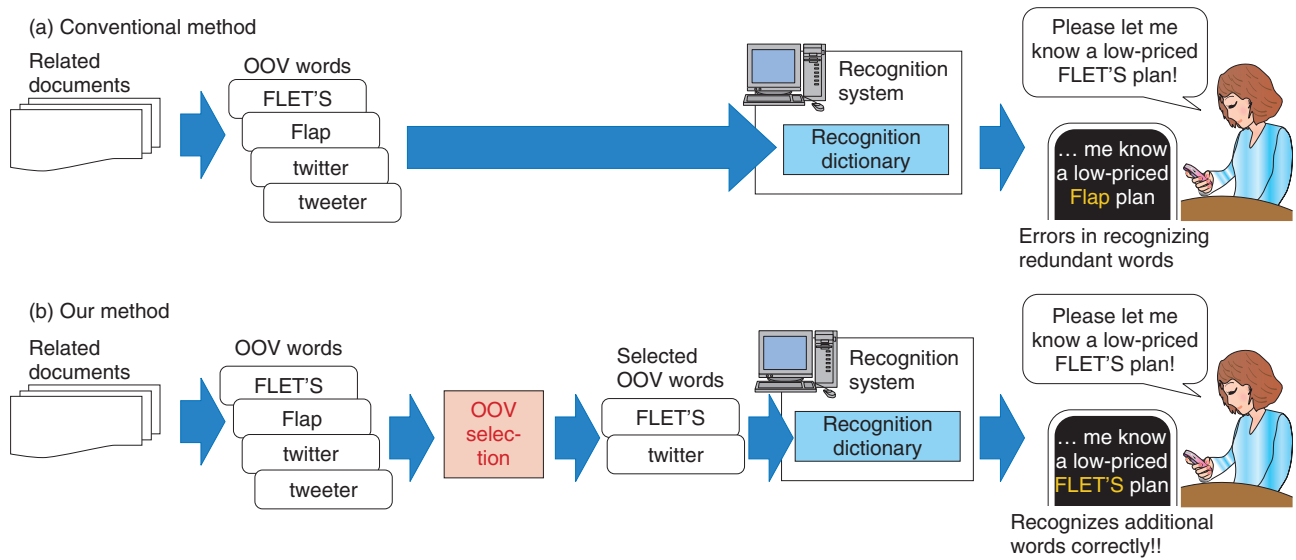


Fig. 5. Comparison of conventional method and our method.

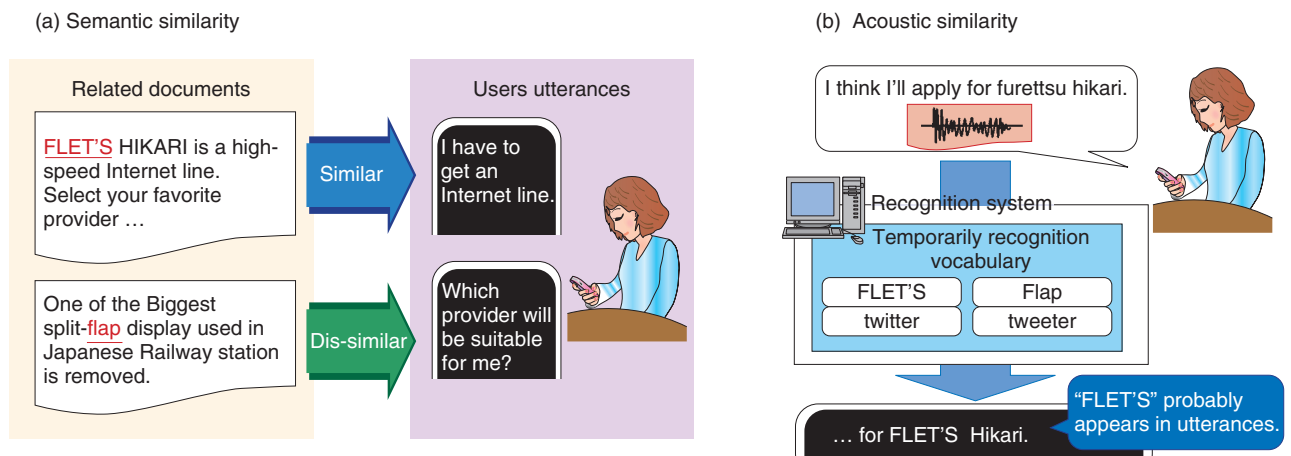


Fig. 6. Mechanism of selecting OOV words using our method.

dropped from the dictionary.

To improve recognition accuracy, we have developed a method that selects only OOV words that will actually be spoken in the user’s utterances (Fig. 5(b)). Our method yields recognition dictionaries that are suitable for each user or service and that can recognize utterances accurately because redundant words are eliminated.

We use two types of information to select OOV words. First, we use the semantic similarity between each OOV word and the user’s utterances (Fig. 6(a)). For example, if a user frequently uttered terms rele-

vant to Internet service such as “Internet”, “line”, and “provider”, we select OOV words that co-occur with these terms such as “FLET’S”. Second, we use acoustic similarity, which refers to whether or not the pronunciation of OOV words is included in the utterances (Fig. 6(b)). For example, if the user uttered “I think I will apply for furettsubikari”, we would assume the word “FLET’S” was probably uttered. To detect the parts of utterances where OOV words appear, we temporarily register all OOV words in the recognition dictionary, and we recognize the utterances using a temporary dictionary to obtain

temporary recognition results. Finally, we select OOV words that appear in the temporary recognition results.

With our method, we were able to reduce the number of recognition errors caused by redundant words by about 10% compared to conventional methods that add all OOV words.

5. Future efforts

We developed the speech recognition engine called VoiceRex to demonstrate the technologies developed at NTT. We plan to implement the technologies introduced in this article in VoiceRex.

We are working on improving the recognition accuracy by enhancing these new technologies to be adapted to each user. We also intend to implement our recognition technologies as cloud services.

References

- [1] Y. Noda, Y. Yamaguchi, K. Ohtsuki, and A. Imamura, "Development of the VoiceRex Speech Recognition Engine," NTT Technical Journal, Vol. 11, No. 12, pp. 14–17, 1999 (in Japanese).
- [2] H. Masataki, D. Shibata, Y. Nakazawa, S. Kobashikawa, A. Ogawa, and K. Ohtsuki, "VoiceRex—Spontaneous Speech Recognition Technology for Contact-center Conversations," NTT Technical Review, Vol. 5, No. 1, pp. 22–27, 2007.
<https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200701022.pdf>
- [3] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A study of mutual front-end processing method based on statistical model for noise robust speech recognition," INTERSPEECH 2009, pp. 1235–1238, Brighton, UK, Sept. 2009.
- [4] T. Asami, N. Nomoto, S. Kobashikawa, Y. Yamaguchi, H. Masataki, and S. Takahashi, "Spoken document confidence estimation using contextual coherence," INTERSPEECH 2011, pp. 1961–1964, Florence, Italy, Aug. 2011.
- [5] S. Yamahata, Y. Yamaguchi, A. Ogawa, H. Masataki, O. Yoshioka, and S. Takahashi, "Automatic Vocabulary Adaptation Based on Semantic Similarity and Speech Recognition Confidence Measure," INTERSPEECH 2012, Portland, OR, USA, Sept. 2012.



Hirokazu Masataki

Senior Research Engineer, Audio Speech and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E., M.E., and D.Eng. degrees from Kyoto University in 1989, 1991, and 1999, respectively. From 1995 to 1998, he worked with ATR Interpreting Telecommunications Research Laboratories, where he specialized in statistical language modeling for large vocabulary continuous speech recognition. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2004 and engaged in the practical use of speech recognition. He received the Maejima Hisoka Award from the Tsushinbunka Association. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Acoustical Society of Japan (ASJ).



Shoko Yamahata

Research Engineer, Audio Speech and Language Media Project, NTT Media Intelligence Laboratories.

She received the B.E. and M.E. degrees from Waseda University, Tokyo, in 2008 and 2010, respectively. She joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2010 and studied language models and vocabulary adaptation. She is a member of ISCA and ASJ.



Taichi Asami

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

He received the B.E. and M.E. degrees from Tokyo Institute of Technology in 2004 and 2006, respectively. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2006 and studied speech recognition, spoken language processing, and speech mining. He received the Awaya Prize Young Researcher Award from ASJ in 2012. He is a member of the International Speech Communication Association (ISCA), IEICE, and ASJ.



Masakiyo Fujimoto

Researcher, NTT Communication Science Laboratories.

He received the B.E., M.E., and D.Eng. degrees from Ryukoku University, Shiga, in 1997, 2001, and 2005, respectively. From 2004 to 2006, he worked with ATR Spoken Language Communication Research Laboratories. He joined NTT in 2006. His current research interests are noise-robust speech recognition including voice activity detection and speech enhancement. He received the Awaya Prize Young Researcher Award from ASJ in 2003, the MVE Award from IEICE SIG MVE in 2008, the Yamashita SIG Research Award from the Information Processing Society of Japan (IPJS) in 2011, and the ISS Distinguished Reviewer Award from IEICE in 2011. He is a member of IEEE, ISCA, IEICE, IPJS, and ASJ.

Knowledge Extraction from Text for Intelligent Responses

*Kugatsu Sadamitsu, Ryuichiro Higashinaka,
Toru Hirano, and Tomoko Izumi*

Abstract

In this article, we describe a linguistic analysis technique for extracting useful knowledge from the vast amounts of text on the World Wide Web. First, we introduce a technique for extracting named entities as a key to knowledge to be handled by a computer, and then we show how this can be applied to extract relations between named entities.

1. Introduction—Text processing aimed at providing an intelligent response

Computers are now able to correctly answer questions such as *How high is Mount Everest?* or to provide information that the user may be interested in but unaware of, such as *It looks like NTT is launching a new service this weekend* via a spoken language interface.

In this article, we introduce the latest techniques to acquire information with a spoken language interface.

For the most part, it seems that the knowledge users require is generally centered around specific entities, such as Mount Everest and NTT in the above examples. Textual expressions relating to such entities are called *named entities*. This means that the expression *Mount Everest* can be uniquely associated with a single entity.

These named entities are crucial to generate intelligent responses. This is because, for example, the named entity *Mount Everest* can be imparted with additional information such as *height* so that an answer can be found for the question *How high is Mount Everest?*, while for the named entity *NTT*, information about the launch of a new service can be gleaned from the World Wide Web (hereafter, the Web) to enable a response in the form, *It looks like NTT is launching a new service this weekend*. If a computer-accessible knowledge database can be con-

structed in this way, then it will become possible to provide intelligent responses (**Fig. 1**).

In this article, we introduce basic techniques for collecting named entities and a technique for extracting the relationships between them.

2. Automatic collection of named entities

If named entities have to be collected to construct a knowledge database, roughly how many named entities are there in the first place? The Wikipedia online encyclopedia contains many named entities; there are over 800,000 entries in the Japanese Wikipedia, and over 4 million in the English Wikipedia. However, these articles are limited to things that are famous or fairly well known, so if we include other less well known people, products, and places, then the total number of entities is quite staggering. Previously, sources of text material were limited to printed media such as newspapers, which imposed severe limitations on the availability of named entities. However, the recent growth of Internet services such as Twitter and blogs has made it possible for users everywhere to publish information by themselves. The Web is now flooded with a wide variety of named entities, and the extraction of these named entities is becoming a very important topic in the construction of knowledge databases.

Since the number of named entities is almost limitless, there is no point trying to manually add each one

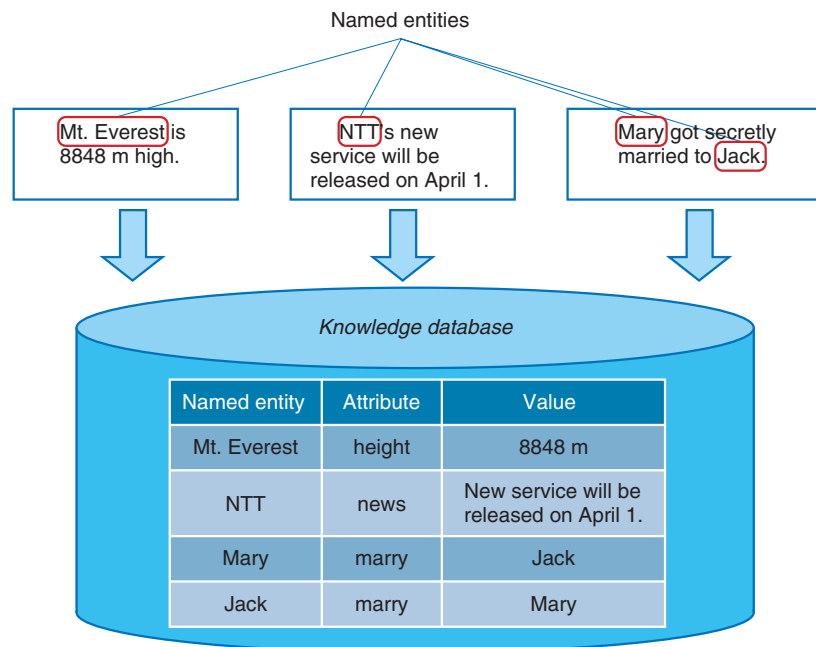


Fig. 1. Knowledge database based on named entities.

to a database. What we need is some way of extracting named entities from text automatically. In recent years, machine learning has become the tool of choice for the extraction of named entities. For example, let's consider the phrase *Today I met Lisa* (Fig. 2). The word *Lisa* in this phrase would normally be considered a named entity that is probably someone's name. This is because we expect *Lisa* to be a person's name based on the surrounding context. In this way, when we identify named entities, we also simultaneously classify them into a category such as the names of people or places. Similarly, machine learning can extract information (specifically, feature quantities) as cues from the surrounding context. On the basis of a statistical model obtained beforehand, we can simultaneously decide whether or not it is a named entity and, if so, which named entity category it belongs to.

Although broad named entity categories such as the names of people or places can be determined from the context, it can be difficult to infer categories with a finer level of detail. Consider the phrase *arrived at K2*. It is understood from the context that *K2* is a place name, but since the phrase provides no additional information, we need other cues in order to achieve a detailed categorization. (*K2* is, in fact, the name of the world's second highest mountain.)

Unlike existing dictionaries where each dictionary tends to have its own category definitions, CGM (consumer-generated media) dictionaries such as Wikipedia offer a high degree of freedom in the assignment of categories. This can cause problems because a systematic categorization tends not to be maintained. Furthermore, the necessary categories are themselves often application-dependent and cannot be used *as-is*.

In the following, we introduce two ways of resolving these issues.

3. Extraction of named entities from text

First, in cases where the characteristics of a named entity cannot be grasped from a single sentence, we considered that it should be possible to grasp the characteristics of the named entity by looking at the document as a whole. For example, if we know that the topic of a document concerns a mountain or a river, then it is possible to characterize the named entity.

So how do we go about grasping the topic of a document? A statistical model called a *topic model* was recently proposed and has been used in many applications [1]. The use of a topic model makes it possible to infer that a document containing words

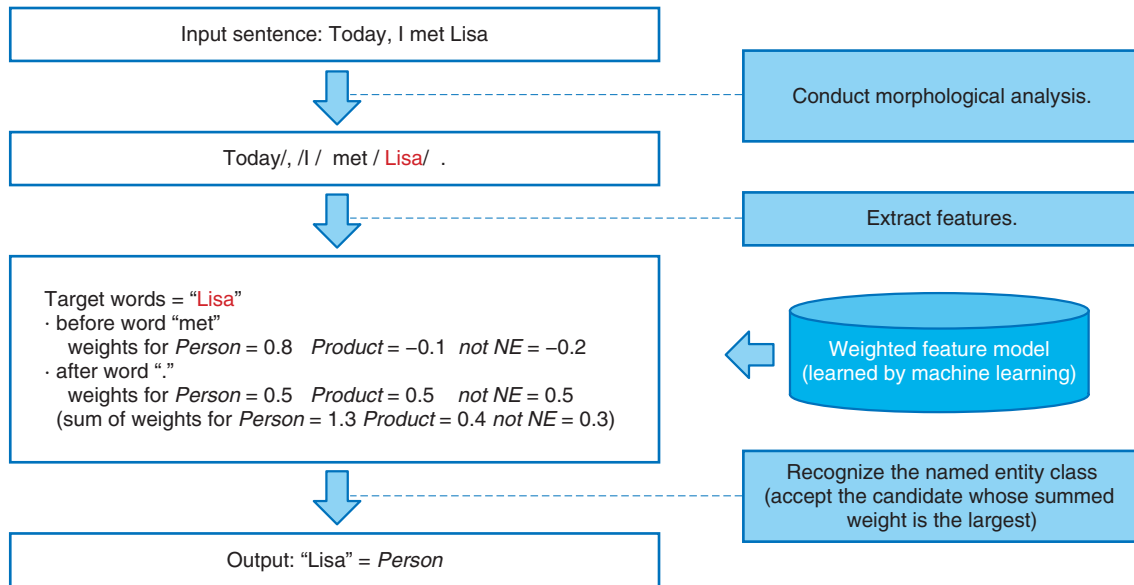


Fig. 2. Automatic extraction of named entities.

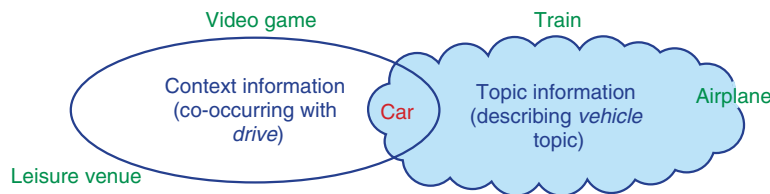


Fig. 3. Illustration of our named entity extraction method, which uses both context information and topic information.

such as *baseball* and *soccer* is probably related to the topic of *sports*.

By combining the global topic information obtained by this topic model with the local context information (and using them as features), we can gather named entities with greater precision [2]. This idea is shown in **Fig. 3**.

For example, suppose we want to gather named entities corresponding to models of vehicles. To extract the names of cars, it is important to look for words such as *drive* in the surrounding context. However, it is difficult to rule out other categories such as holiday destinations or motor racing video games without additional cues, since these categories can also co-occur with the word *drive*. Even if we assume it has been established that the topic is *vehicles*, there is still some ambiguity with the topic information

alone, which could refer to *trains* or *airplanes*. Only when it is used in conjunction with context information can the ambiguity be greatly reduced, making it possible to accurately infer the correct category of car names.

4. Creating a named entity dictionary from existing dictionaries

The second method involves making use of existing dictionaries. As discussed above, existing dictionaries each have their own category definitions but do not share a unified categorization system. If these categories could be mapped to our desired named entity categories, then it would be possible to use existing dictionaries for the automatic extraction of named entities.

It hardly needs to be said that existing dictionaries are treasure troves of information. For example, if the Wikipedia entry for *K2* is assigned to the category *Himalayas*, then this suggests that it maps to a named entity in the *names of mountains* category. The phrase *the mountain in ~* at the head of the description also provides a useful clue. We have developed a technique that can perform accurate category mapping by combining machine learning with clues obtained from multiple viewpoints in this way [3].

Although this technique has made it possible to extract named entities with high precision, it still has a weak point in that its applicable range is limited to well-known named entities. Therefore, a challenge for the future is to develop a better technique that can be used in conjunction with the automatic extraction of named entities from text as described above.

5. Extraction of relationships between named entities

So far, we have introduced a method that automatically constructs a named entity dictionary labeled with category information. However, a named entity dictionary on its own is not able to answer complex questions such as *Who did Maria marry?* To be able to do this sort of advanced processing, we need new information associated with the named entity. As an example of a practical technique that deals with named entities, we considered a method that extracts relationships between named entities from text. For example, let's consider how relationships between named entities can be extracted from the sentence *Nancy was shocked that Maria got secretly married to Jack.* We'll assume that the named entities *Maria*, *Nancy* and *Jack* are extracted. This sentence says that *Maria* and *Jack* have the relationship *married*, and that *Nancy* and *Jack* have no relationship. If we simply consider the surrounding named entities to be related, then we would mistakenly extract a relationship between *Maria* and *Nancy*, even though nothing is said about the relationship between these named entities.

Therefore, as shown in **Fig. 4**, we have developed a technique based on the results of dependency analysis that uses cues derived from the relative positioning of named entities to figure out if these entities are related, and if so, how they are related [4]. For example, we can see that the clauses about *Maria* and *Jack* are both connected to the *married* clause. We can therefore conclude that *Maria* and *Jack* have the relationship *married*. More accurate extraction of relation-



Fig. 4. Extraction of relationships between named entities.

ships between named entities can be achieved by combining the results of dependency analysis with a method that automatically identifies whether or not the surrounding words indicate a lexical relationship between the named entities based on a large-scale text corpus.

6. Future work

In this article, we have introduced techniques for extracting knowledge from text in order to generate intelligent responses, with a particular focus on named entities. These techniques can be used in many different applications, including technology that can answer questions like the one posed in the Feature Articles entitled “Question Answering Technology for Pinpointing Answers to a Wide Range of Questions” [5]. It can also be used in search engines and the like.

Targets of knowledge extraction are not only named entities, but also relationship information (as described above), reputation information, and information to infer the attributes of blog/Twitter users. We will continue to further our study of knowledge extraction in the future in order to address the increasingly diverse needs of society and to propose new services.

References

- [1] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.
- [2] K. Sadamitsu, K. Saito, K. Imamura, and G. Kikui, “Entity Set Expansion Using Topic Information,” *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-HLT 2011): Human Language Technologies*, pp. 726–731, Portland, OR, USA.
- [3] R. Higashinaka, K. Sadamitsu, K. Saito, T. Makino, and Y. Matsuo, “Creating an Extended Named Entity Dictionary from Wikipedia,” *Proc. of the 24th International Conference on Computational Linguistics (COLING 2012)*, pp. 1163–1178, Mumbai, India.
- [4] T. Hirano, H. Asano, Y. Matsuo, and G. Kikui, “Recognizing Relation Expression between Named Entities Based on Inherent and Context-dependent Features of Relational Words,” *Proc. of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pp. 409–417, Beijing, China.
- [5] R. Higashinaka, K. Sadamitsu, K. Saito, and N. Kobayashi, “Question Answering Technology for Pinpointing Answers to a Wide Range of

Questions,” NTT Technical Review, Vol. 11, No. 7, 2013.
<https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa4.html>



Kugatsu Sadamitsu

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

He received the B.E., M.E., and Ph.D. degrees in engineering from Tsukuba University, Ibaraki, 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).



Toru Hirano

Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E. degree in systems engineering from Wakayama University and the M.E. and Ph.D. degrees in information science from Nara Institute of Science and Technology in 2003, 2005, and 2012, respectively. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2005. His current research interests include information extraction and user profile inference. He is a member of NLP.



Ryuichiro Higashinaka

Senior Research Scientist, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.A. degree in environmental information, the Master of Media and Governance, and the Ph.D. degree 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 Nov. 2004 to Mar. 2006, he was a visiting researcher at the University of Sheffield in the UK. From 2006 to 2008, he was a part-time lecturer at Osaka Electro-Communication University. Since 2010, he has been a part-time lecturer at Keio University. He is a member of the Japanese Society for Artificial Intelligence, IPSJ, and NLP.



Tomoko Izumi

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

She received the B.A. degree in global education from Hokkaido University of Education, in 2005, the M.A. degree in applied linguistics from Boston University, Massachusetts, USA, in 2007, and the M.E. degree in English Language Education from Hokkaido University of Education in 2008. She joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2008. Her current research focuses on automatic recognition of synonyms. She is a member of NLP.

Question Answering Technology for Pinpointing Answers to a Wide Range of Questions

Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, and Nozomi Kobayashi

Abstract

Question answering technology that provides pinpoint answers to a wide range of questions is expected to make speech interfaces more intelligent. This article describes question answering technology and reviews past and current approaches to it at NTT.

1. Introduction

The sheer number of documents that have been created on the Internet makes it impossible for one to read them all. Therefore, technologies that help us find pertinent information efficiently are becoming more and more important. One such technology is the web search engine. Web search engines search for documents in response to keywords provided by users, present the documents, and thereby facilitate our information access. Although web search engines are useful, they only return documents; that is, users still need to read through the returned documents for the information they need. It is easy to imagine that the amount of information we handle will increase at a pace faster than ever before, making it too time-consuming to even look through the returned documents. There is therefore an imminent need for technologies that make our information access more efficient. In this article, we describe question answering technology, which represents a major step forward in meeting this need. We also describe NTT's past and current approaches to question answering technologies.

2. Question answering technology

Question answering technology provides pinpoint answers to natural language questions. Systems based on this technology are called *question answering sys-*

tems. Users can obtain pinpoint answers to their questions just by asking the system; answers are presented immediately, and there is no need to read any documents. Question answering systems deal with a wide variety of questions. There are two types of systems depending on the type of questions they answer. One is a factoid question answering system, which answers factual questions by using words or short phrases. The other is a non-factoid question answering system, which provides answers in sentence or paragraph form. For example, the former system answers *Everest* to *What is the highest mountain in the world?* and *George Washington* to *Who was the first president of the United States?* The latter, for example, provides sentential answers to definition questions (e.g., *What is optical fiber?*), why-questions (e.g., *Why is the sky blue?*), and how-questions (e.g., *How do I cook delicious dumplings?*). If all questions that could be asked by users were known in advance, it would be possible to prepare the answers in advance and provide the correct answers when needed. However, in reality, user questions are too diverse to predict. Therefore, question answering systems mimic how humans find answers; that is, they interpret a question, search for relevant documents, and find answers.

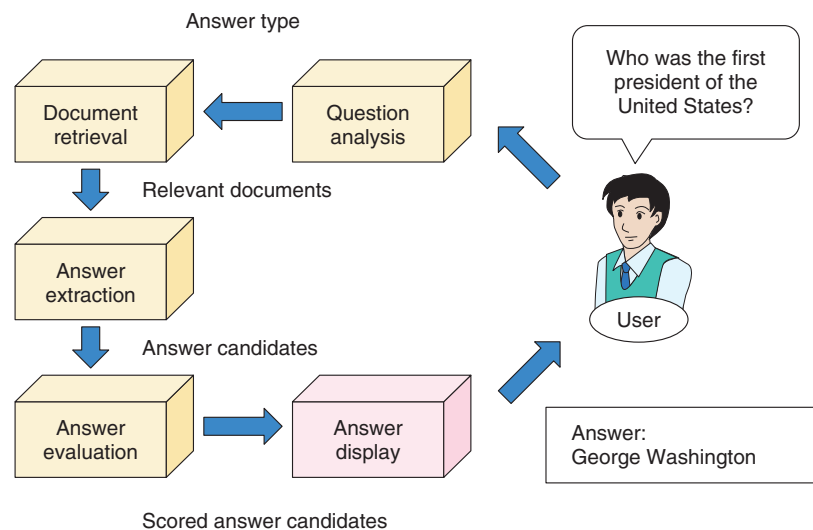


Fig. 1. General architecture of a question answering system.

3. Architecture of a question answering system

The general architecture of a question answering system is shown in Fig. 1. It has four modules: question analysis, document retrieval, answer extraction, and answer evaluation [1]. Note that although this article deals with a factoid question answering system, the architecture is almost the same for a non-factoid question answering system.

First, the question analysis module analyzes a question and determines its answer type. An answer type represents the type of information requested by a question: a person's name, place name, and a numerical expression are some of the possible answer types. For *Who was the first president of the United States?*, the answer type is *person* (person's name). When the granularity of answer types becomes fine-grained, it is possible to grasp the requested information with pinpoint precision, although there is a trade-off between the number of answer types and the accuracy of automatic answer-type classification. The information retrieval module uses a search engine to retrieve relevant documents on the basis of keywords in the question. Since the question answering system searches for answers only in those retrieved documents, the accuracy of document retrieval is very important. The answer extraction module extracts from the retrieved documents all answer candidates matching the answer type. When the answer type is *person*, all person names in the retrieved documents are extracted. Named entity recognition technology

[2] is used for this extraction. Finally, the answer evaluation module evaluates the appropriateness of candidate answers by using such information as how they appear in the documents, and assigns scores to the candidate answers. Finally, highly scoring candidate answers are presented to the user.

4. Question answering systems at NTT

Research on question answering began around 1999. Around that time, an evaluation workshop on question answering was held at the Text Retrieval Conference (TREC), and researchers from all over the world started competing with one another to create question answering systems. At around the same time, NTT also started researching question answering systems and since then has developed a number of systems.

The first question answering system developed at NTT was SAIQA (System for Advanced Interactive Question Answering) in 2001 [3]. SAIQA was a factoid question answering system that obtained accurate answers thanks to its early use of machine learning techniques for answer-type classification and named entity recognition. Machine learning is a process of statistically learning criteria for judgment from a large amount of training data. In 2004, SAIQA achieved the best performance at the evaluation workshop Question Answering Challenge (QAC) in Japan. SAIQA used a database of newspaper articles for document retrieval.

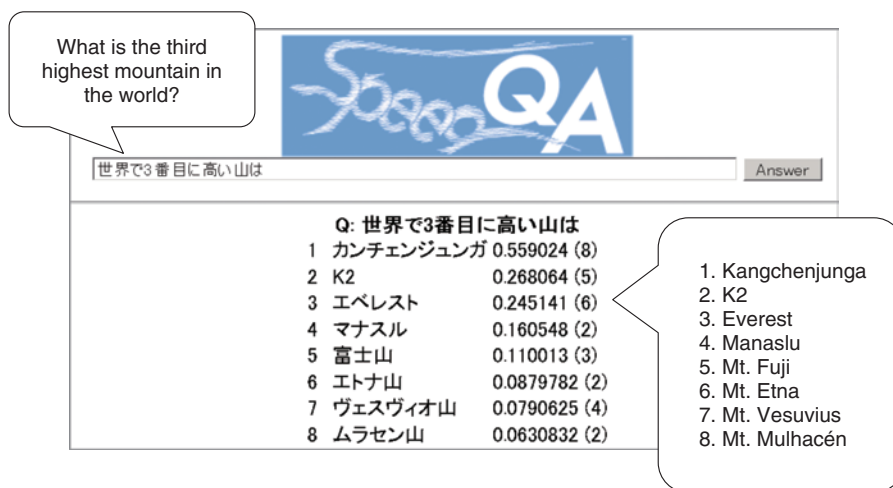


Fig. 2. Screenshot of SpeedQA. The numbers indicate scores of candidate answers and their frequencies (in parentheses) in retrieved documents.

In 2004, Web Answers was developed [1]. This system, as its name implies, retrieves documents from the web. The system was made public on the Internet so that anyone could use it. It answered not only factoid questions but also definition and reputation questions (e.g., *What is the reputation of X?*), which made it popular among users for its wide coverage. In 2007, we developed a non-factoid question answering system called NAZEQA in order to provide answers to why-questions [4]. Conventional approaches to why-questions had used cue words such as *node* and *tame* (corresponding to *because* in English) to find answer sentences. However, it was also pointed out in the literature that relying only on such cue words resulted in limited coverage of answers. Therefore, we statistically mined causal expressions from a large number of documents and used them to detect answer sentences, which achieved accurate answers to why-questions. In 2012, we developed SpeedQA, our most recent system. A screenshot is shown in **Fig. 2**. This system is the culmination of our experience in researching question answering systems at NTT. For example, it uses machine learning in almost all of its modules. It uses the Internet for document retrieval and can answer factoid as well as non-factoid questions. Non-factoid questions it can deal with are definition, reputation, and also why-questions. This system, minus some of its functions, has been integrated into the knowledge Q&A service of NTT DOCOMO's Shabette-Concier* voice-agent service [5].

5. Recent advances: Answer-type classification and timeliness detection in SpeedQA

5.1 Answer-type classification

In developing SpeedQA, we concentrated in particular on answer-type classification because we wanted to provide pinpoint answers. The answer analysis module of SpeedQA first determines whether a question type is factoid or non-factoid. Then, for a factoid question, it further classifies its answer type by referring to a taxonomy of over 100 answer types. Even with machine learning, classification into such fine-grained answer types is not an easy task. Therefore, we used a large-scale Japanese thesaurus created at NTT [6] and put a special focus on noun suffixes and counter suffixes. In this way, we succeeded in finding useful information for answer-type classification and achieved high accuracy.

5.2 Timeliness detection

To find pinpoint answers, in addition to accurately classifying answer types, it is also important to recognize user intentions, that is, to determine what users really want to know. Consider the question *Who won the gold medal?* On the surface, this question looks like an easy one for which the system can simply present past gold medalists. However, if it is posed during the Olympic Games, it would be reasonable to present only the gold medalists in the current

* Shabette-Concier is a trademark of NTT DOCOMO Inc.

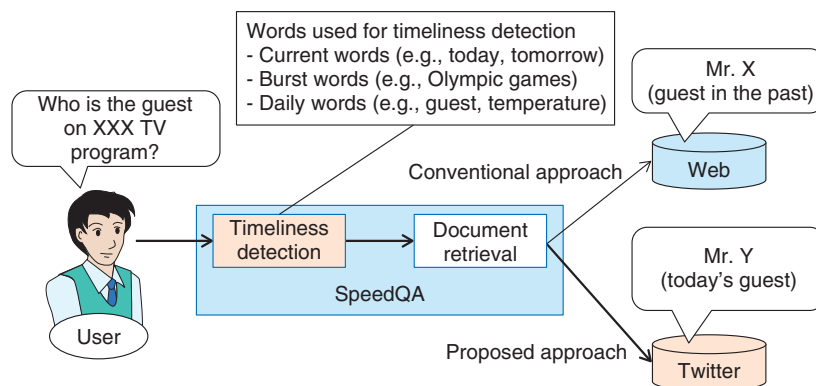


Fig. 3. Flow of timeliness detection.

Olympics. As shown by this example, there are questions whose answers vary depending on when they are posed, and they need to be treated accordingly to provide pinpoint answers. To this end, we developed a method of timeliness detection that detects whether a question is asking about timely events or not. If it is detected as being timely, the document retrieval module is switched from a web search engine to a real-time search engine (a search engine for Twitter) for timely information. The flow of timeliness detection is shown in **Fig. 3**. When questions contain time-related words such as *today* or *now* (called *current words*), it is easy to detect their timeliness. When questions contain *burst words* (words whose occurrence frequencies have shown a sudden increase in a short time span), it is also reasonable to determine that they are timely questions. The problem is when a question does not contain such words. We collected and analyzed many questions that asked for timely information and discovered that words such as *guest*, *starting player*, *game*, and *temperature*, whose referents (entities being referred to) vary day by day, are used frequently. We named such words *daily words* and developed a technique for automatically acquiring them. By using this technique, we were able to obtain many daily words, and it became possible to successfully detect the timeliness of questions containing such words.

6. Conclusion

This article described question answering technology and past and current approaches to this technology at NTT. We acknowledge that our latest system SpeedQA still has a lot of room for improvement. We plan to improve our algorithms further to achieve more pinpoint answers.

References

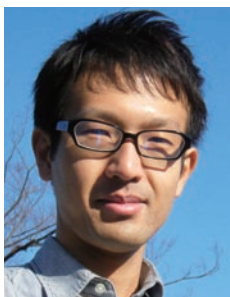
- [1] H. Isozaki, R. Higashinaka, M. Nagata, and T. Kato, "Question Answering Systems," Corona Publishing Co. Ltd., 2009 (in Japanese).
- [2] K. Sadamitsu, R. Higashinaka, T. Hirano, and T. Izumi, "Knowledge Extraction from Text for Intelligent Responses," *NTT Technical Review*, Vol. 11, No. 7, 2013. <https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa3.html>
- [3] E. Maeda, H. Isozaki, Y. Sasaki, H. Kazawa, T. Hirao, and J. Suzuki, "Question answering system: SAIQA—A "Learned Computer" that answers any questions," *NTT R&D*, Vol. 52, No. 2, pp. 122–133, 2003 (in Japanese).
- [4] R. Higashinaka and H. Isozaki, "Automatically acquiring causal expression patterns from relation-annotated corpora to improve question answering for why-questions," *ACM Transactions on Asian Language Information Processing*, Vol. 7, No. 2, 2008.
- [5] W. Uchida, C. Morita, and T. Yoshimura, "Knowledge Q&A: Direct Answers to Natural Questions," *NTT DOCOMO Technical Journal*, Vol. 14, No. 4, pp. 4–9, 2013. http://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/rd/technical_journal/new/vol14_4_004en.pdf
- [6] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi, "Goi-Taikai—A Japanese Lexicon," Iwanami Shoten, 1997.



Ryuichiro Higashinaka

Senior Research Scientist, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

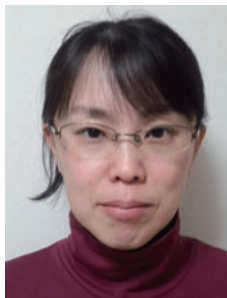
He received the B.A. degree in environmental information, the Master of Media and Governance, and the Ph.D. degree from Keio University, Kanagawa, in 1999, 2001, and 2008, respectively. He joined NTT in 2001. His research interests include building question answering systems and spoken dialogue systems. From Nov. 2004 to Mar. 2006, he was a visiting researcher at the University of Sheffield in the UK. From 2006 to 2008, he was a part-time lecturer at Osaka Electro-Communication University. Since 2010, he has been a part-time lecturer at Keio University. He is a member of the Japanese Society for Artificial Intelligence, the Information Processing Society of Japan (IPSJ), and the Association for Natural Language Processing (NLP).



Kugatsu Sadamitsu

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

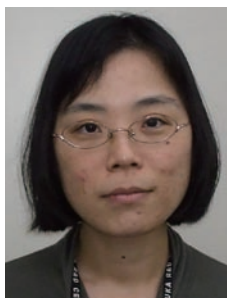
He received the B.E., M.E., and Ph.D. degrees in engineering from Tsukuba University, Ibaraki, in 2004, 2006, and 2009, respectively. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2009. His current research interests include natural language processing and machine learning. He is a member of IPSJ and NLP.



Kuniko Saito

Senior Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

She received the B.S. and M.S. degrees in chemistry from the University of Tokyo in 1996 and 1998, respectively. She joined NTT Information and Communication Systems Laboratories in 1998 and then moved to NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories). Her research focuses on part-of-speech tagging, named entity recognition, and term extraction. She is a member of IPSJ and NLP.



Nozomi Kobayashi

Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

She received the M.E. and Dr.Eng. degrees in information science from Nara Institute of Science and Technology in 2004 and 2007, respectively. She joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2007. Her current research interests include information extraction. She is a member of IPSJ and NLP.

Speech Synthesis Technology to Produce Diverse and Expressive Speech

Hideyuki Mizuno, Hideharu Nakajima, Yusuke Ijima, Hosana Kamiyama, and Hiroko Muto

Abstract

We have been developing a new text-to-speech synthesis system based on *user-design* speech synthesis technology that can be extensively applied to various fields. The technology yields speech with rich expression and various characteristics and thus replaces existing synthesized speech systems that have a limited range of voices or speaking styles. This article introduces this new system that represents the future of speech synthesis technology.

1. Introduction

The use of mobile phone telecommunication services is continuing to increase, and this is driving demand for various speech synthesis services, for example, speech guidance and speech dialogue services. For such services, speech synthesis must offer not only high quality but also variety. For example, synthesized speech that remains audible even in noisy environments and that has a characteristic voice quality and speaking style is required. The Cralinet (CReate A LIkeNEss to a Target speaker) system originally developed by the NTT Media Intelligence Laboratories as a telephone speech guidance service can generate high quality speech [1]. Cralinet has been broadly used in a safety confirmation system and in an automatic speech guidance system for business contact centers. The main feature of Cralinet is the production of synthesized speech that is as natural as that of humans. This was achieved by using a lot of speech waveforms uttered by a narrator and properly connecting them. Unfortunately, only one voice, that of a female speaker, is output, and the speaking style is limited to reading. Hereafter, the main aim of our research and development activities will be to introduce various new speech services that satisfy a far wider range of demands. The immediate goals are to

generate any kind of voice or style while retaining the usability of speech even in noisy environments. In this article, we introduce the *user design* speech synthesis technology, which can generate expressive synthesized speech.

2. Outline of user-design speech synthesis technology

The framework of our technology is shown in **Fig. 1**. First, when the target speaker's voice is input, a source model is selected according to the voice quality of the speaker. The source model is then trained using the acoustic features of the voice. Next, the texts given as the speech synthesis target are analyzed to determine the most appropriate speaking style, e.g., a reading, storytelling, or sales pitch style. The resulting synthesized speech thus has the voice quality of the target speaker and also the appropriate speaking style. If the end-use environment is discovered to be noisy, the speech is enhanced to permit clear discernment.

The *user-design* speech synthesis technology comprises three novel techniques: automatic model training, speaking style assignment, and speech clarity enhancement. Some conventional text-to-speech technologies that provide similar functions have been

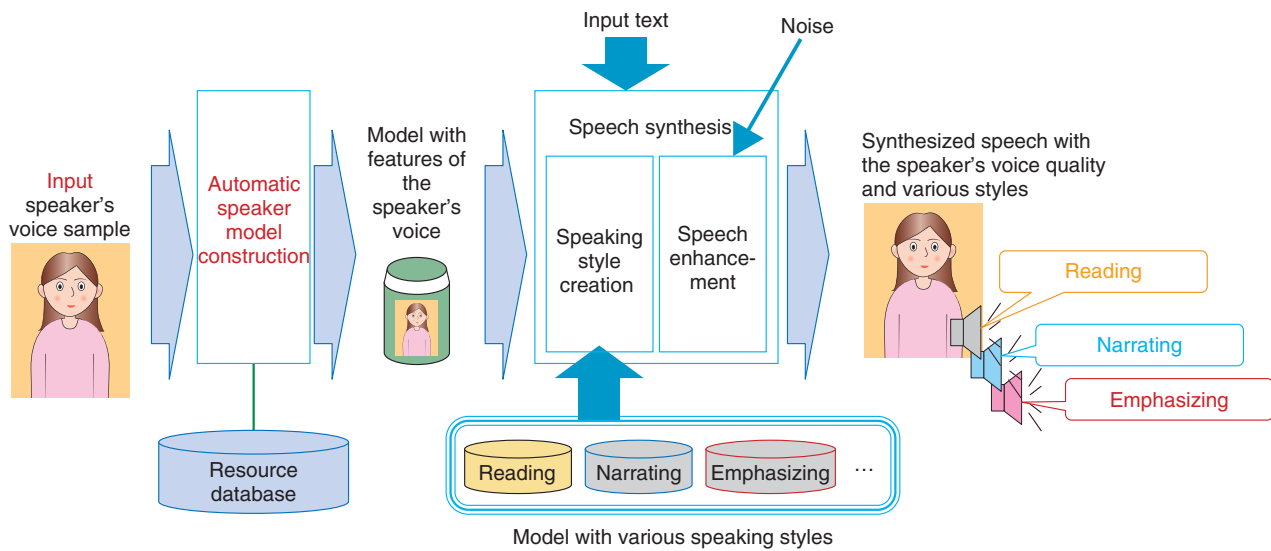
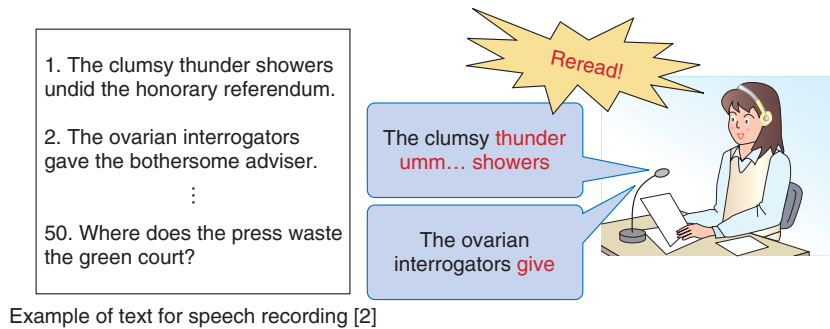


Fig. 1. Framework of user-design speech synthesis technology.



Example of text for speech recording [2]

Fig. 2. Example of the speech recording process for speech synthesis.

developed, but they have several problems. For instance, the time required to train the speaker model is excessive, speaking styles are limited, and speech clarity enhancement is effective only for a specific kind of noise. We apply our three new techniques to produce synthetic speech with rich expression and various characteristics and to realize speech synthesis with a voice similar to a user-specified speaker's voice based on very little speech data from the target speaker. Moreover, we can generate prosody, which refers to the rhythm, stress, and intonation of speech, in order to produce speech with various speaking styles, and we can enhance speech by using noise characteristics to set speech features and thus maintain high voice quality.

3. Synthesis of various speakers

Recent advances in speech synthesis technology mean that it is now possible to achieve reading out of various texts in a specified speaker's synthesized speech, but only if about one hour of speech data is uttered by that speaker. Moreover, as shown in Fig. 2, the desired speaker must utter the set text precisely word-for-word. Reading errors are common, so the same text must be reread until the samples are error-free. This is not a problem for professional narrators, but it is impractical for the general public (family members or friends). Clearly, the amount of recorded speech data required must be reduced. Our solution is called arbitrary speakers' speech synthesis. This solution can synthesize speech that sounds like any

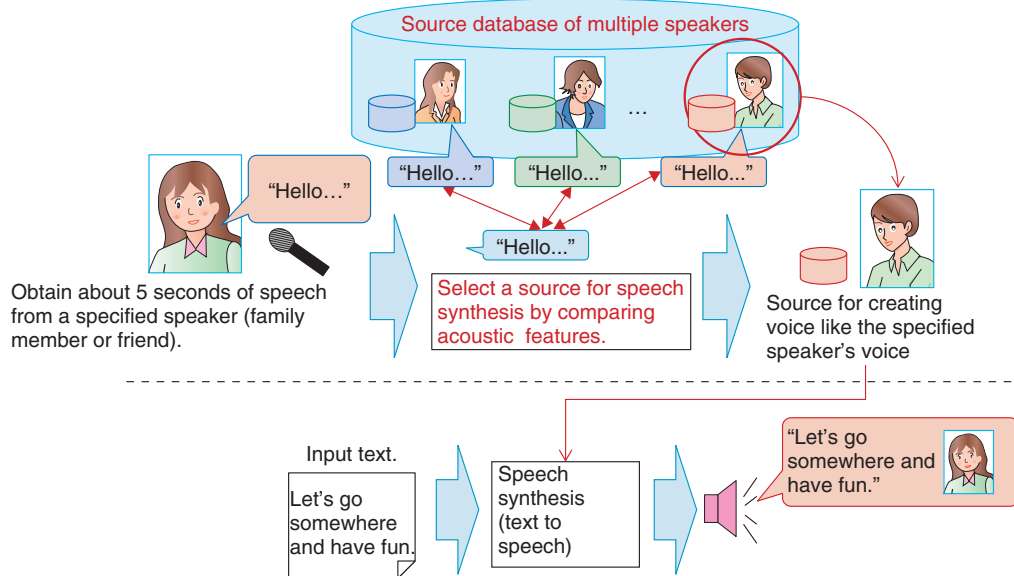


Fig. 3. Overview of arbitrary speakers' speech synthesis technique.

particular speaker from just 5 sec of the speaker's speech data.

An overview of the technique is shown in **Fig. 3**. Samples of speech from multiple speakers are obtained in advance. These become the sound sources for speech synthesis that form the source database. By obtaining just a very small amount of a specified speaker's speech data, i.e., a speaker not in the database, and comparing the acoustic features of that data with the previously obtained samples, a sound source can be selected from the source database to create a voice very similar to the specified speaker's voice. This technique therefore makes it possible to synthesize speech that sounds like any particular speaker based on only 5 sec of the speaker's speech data. Experiments confirmed that 70% of the speakers selected from the database using this technique had a similar voice to the target speaker's voice.

4. Expressive speaking style

The style of speech depends on where the speech is uttered and what the intended purpose is, so adding a natural speaking style for various domains yields expressive synthesized speech. This kind of synthesis research is known as *expressive speech synthesis* and is being researched worldwide. Style can be expressed by three factors: i) intonation, ii) speed, and iii) loudness of speech. The style is determined by combining

these three factors. Of these factors, intonation is known to be the most perceptible factor.

We recorded both conventional reading style speech and expressive style speech, and compared their intonations. We observed that 1) expressive speech had higher intonation than reading speech in many phrases when a phrase-by-phrase comparison of fundamental frequency (F0) was done, and 2) there are various F0 movements at phrase-end positions, for example, rise, fall, rise-fall, and rise-fall-rise. The first observation is described as *phrase emphasis*, and an example of higher F0 is shown in **Fig. 4(a)**. The second is called the *phrase boundary tone*. As shown in **Fig. 4(b)**, although the phrase boundary tone in reading style speech falls towards the phrase end, e.g., "Tsukare-masen-yo↘" (in English, "You may not be tired"), the tone in expressive style speech rises around the end of the phrase, e.g., "Tsukare-masen-yo↗" to strongly emphasize the message.

These two phenomena of *phrase emphasis* and *phrase boundary tone* are found to be useful as F0 generation control factors when synthesizing speech with the hidden Markov model (HMM), which is commonly used in many studies [3]. To achieve expressive text-to-speech synthesis (which takes text as input and generates synthesized speech as output), we investigated a method for predicting whether or not the phrase boundary tone rises at each phrase end [4]. For this prediction, it is not sufficient to know the

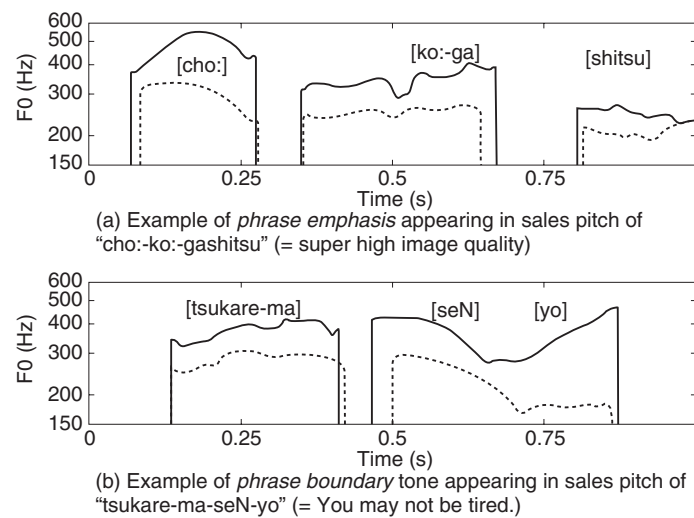


Fig. 4. Fundamental frequency (F0) difference between reading and expressive style speech. Solid line: expressive, dashed line: reading.

identities of the phrase-end particles. Phrase boundary tones change with the context surrounding the phrase boundary and the situation in which the speech is uttered e.g., “A-ka \uparrow (asking question)” vs. “A-ka \downarrow (with disappointment)”. Though some phrase boundary rise/fall prediction rules can be written by human experts, the variation is too large due to speaker individuality and the diversity of domains and situations. Thus, we use speech/linguistic data and a machine learning method to construct models to predict phrase boundary rise/fall. Through experiments targeting expressive speech such as sales pitches and telephone call center operation conversations, we confirmed that the proposed method can accurately predict phrase boundary rise/fall labels.

5. Enhanced synthesized speech

To apply synthesized speech to a wide variety of speech services, the synthesized speech must not only be expressive but also intelligible. In noisy environments, speech can be hard to follow unless some form of noise cancellation is used. Hence, we have been developing a technique to enhance synthesized speech. It yields synthesized speech that remains discernible while retaining as much of the distinctive characteristics of a speaker’s voice and speaking style as possible. As the first step, we analyzed the attributes of easily discerned speech in noisy environments. It is well known that some speakers have

voices that carry exceptionally well; they cut through noise and are easily heard.

We investigated the attributes of such *carrying voices* using many speakers and several types of noise. The experiments revealed that the carrying voices had a higher power spectrum in specific frequency bands occupied only by vowel sounds (which are produced by vibrating the vocal cords) than the noise [5]. As the next step, we developed a technique that uses the results of analysis to reproduce the carrying voice without changing the unique characteristics of the speaker’s voice. This is done by accentuating the power spectra of the specific frequency bands so that they dominate the identical frequency bands of the noise.

The direct enhancement of power causes unexpected and unwanted changes in voice quality. Therefore, our enhancement algorithm first identifies the vowel parts of the synthesized speech from pronunciation information generated in the speech synthesis process. Next, the power spectra of the frequency bands are increased. It is difficult to precisely determine the frequency bands from actual speech in real time because the bands vary with the pronunciation. However, with speech synthesis they can be accurately determined prior to their use by analyzing the speech source. Therefore, synthesized speech that is both intelligible and high in quality can be achieved by using both pronunciation information and the frequency characteristics at specific bands, as shown in

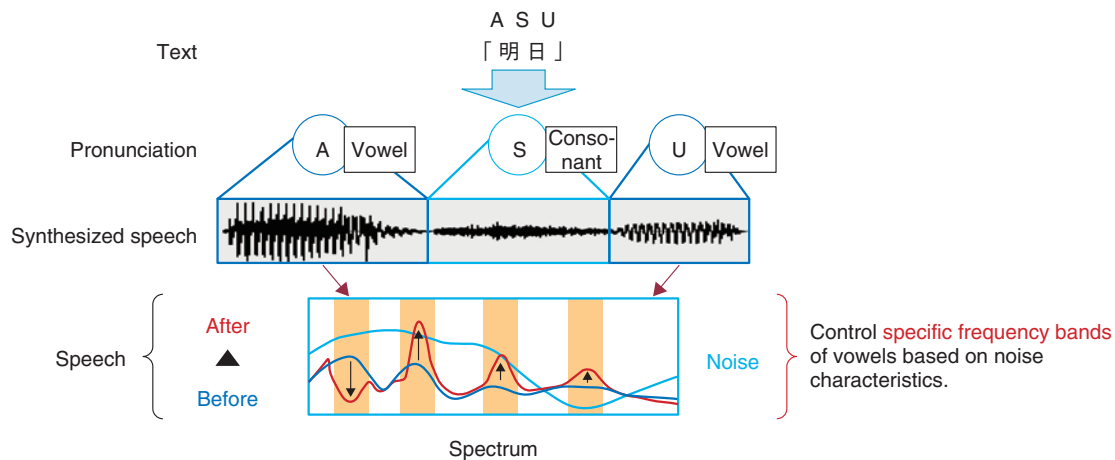


Fig. 5. Process of clarifying speech synthesis result.

Fig. 5. The result is conducive to an increase from 50–60% to 80% in word discernment. This result indicates that the technique significantly increases the appeal and utility of synthesized speech.

6. Conclusion

The techniques introduced in this article are able to yield expressive synthetic speech with high voice quality and various speaking styles and that offers excellent clarity even in noisy environments. With these techniques, the range of applications of speech synthesis will expand greatly from the conventional applications, which have been restricted by the limited variety of speech and use environments. When the expressive speech synthesis technique described here is refined, the usage of speech synthesis will expand to encompass speech dialogue systems that can talk with various voice qualities and speaking styles in accordance with user requests. Optimizing the techniques introduced in this article is our im-

mediate goal.

References

- [1] 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, 2007. <https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200701028.pdf>
- [2] A. W. Black and K. Tokuda, “The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets,” *Proc. of Interspeech 2005*, pp. 77–80, Lisbon, Portugal, 2005.
- [3] Y. Maeno, T. Nose, T. Kobayashi, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka, “A study on prosodic contextual factors for HMM-based speech synthesis with diverse speaking styles,” *Proc. of the Acoustic Society of Japan 2011 Spring Meeting*, pp. 385–386, 2011 (in Japanese).
- [4] H. Nakajima and H. Mizuno, “Predicting phrase boundary tone labels for expressive text-to-speech synthesis,” *Proc. of the Acoustic Society of Japan Autumn Meeting*, pp. 361–362, 2011 (in Japanese).
- [5] H. Kamiyama, Y. Ijima, M. Isogai, and H. Mizuno, “Analysis of the correlation between various acoustic features and the audibility of speech with noise,” *IEICE Technical Report*, Vol. 122, No. 81, pp. 69–74, 2012 (in Japanese).



Hideyuki Mizuno

Senior Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E. and M.E. degrees from Nagoya University, Aichi, in 1986 and 1988 and the Dr.Eng. degree in systems and information engineering from the University of Tsukuba, Ibaraki, in 2006. In 1988, he joined NTT Human Interface Laboratories, where he engaged in R&D of speech synthesis and voice quality conversion. During 1994–1997, he worked for NTT Intelligent Technology Co. Ltd. developing speech application systems. He is currently researching text-to-speech synthesis and its applications. Since 2009, he has been an Associate Editor of the Editorial Board of the Acoustic Society of Japan (ASJ). Since 2010, he has been a chair of the Speech Synthesis Group of the Technical Standardization Committee on Speech Input/Output Systems in the Japan Electronics and Information Technology Industries Association, Japan. He received the Technical Development Award from ASJ in 1998. He is a member of ASJ and the Institute of Electronics, Information and Communication Engineers (IEICE).



Hosana Kamiyama

NTT EAST.

He received the B.E. and M.E. degrees in computer science from Tokyo Institute of Technology in 2008 and 2010, respectively. Since joining NTT in 2010, he had been engaged in researching synthesized speech enhancement. He is a member of ASJ and IEICE. He moved to NTT EAST in July 2013. At the time of this research was conducted, he was Researcher, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.



Hiroko Muto

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

She received the B.E. and M.E. degrees in computer science from Tokyo Institute of Technology in 2009 and 2011, respectively. Since joining NTT in 2011, she has been engaged in researching text processing for speech synthesis. She is a member of ASJ.



Hideharu Nakajima

Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E. and M.E. degrees in computer engineering and information science from the University of Tokushima in 1990 and 1992 and the Ph.D. degree in global information and telecommunication studies from Waseda University, Tokyo, in 2010. He joined NTT Information Processing Laboratories in 1992. During 1997–2002, he worked for Advanced Telecommunications Research Institute International (ATR). His research interests include spoken/natural language processing based on clear principles and he is now investigating corpus-based text-processing for speech synthesis. He is a member of ASJ, the Phonetic Society of Japan, the Association for Natural Language Processing (NLP), IEICE, the Information Processing Society of Japan, and the Japanese Cognitive Science Society.



Yusuke Ijima

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

He received the B.E. degree in electrical and electronics engineering from the National Institution for Academic Degrees and University Evaluation after graduating from Yatsushiro National College of Technology, Kumamoto, in 2007, and the M.E. degree in information processing from Tokyo Institute of Technology in 2009. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2009 and began researching speech synthesis. His research interests include speech synthesis, speech recognition, and speech analysis. He is a member of ASJ and the International Speech Communication Association.

Multichannel Audio Transmission over IP Network by MPEG-4 ALS and Audio Rate Oriented Adaptive Bit-rate Video Codec

Yutaka Kamamoto, Noboru Harada, Takehiro Moriya, Sunyong Kim, Takahiro Yamaguchi, Masanori Ogawara, and Tatsuya Fujii

Abstract

This article describes an experiment of lossless audio transmission over an Internet protocol (IP) network and introduces a prototype codec that combines lossless audio coding and variable bit rate video coding. In the experiment, 16-channel acoustic signals compressed by lossless audio coding (MPEG-4 Audio lossless coding (ALS) standardized by the Moving Picture Experts Group) were transmitted from a live venue to a café via the IP network. At the café, received sound data were decoded losslessly and then appropriately remixed to adjust to the environment at that location. The combination of high-definition video and audio data enables fans to enjoy a live musical performance in places other than the live venue. This experiment motivated us to develop a new prototype codec that guarantees high audio quality. The developed codec can control the bit rates of both audio and video signals jointly, and it achieves high audio and video quality.

1. Introduction

Network quality has improved recently, and the storage size has also been rapidly expanding. The Next Generation Network (NGN) can provide high quality, secure, and reliable services [1]. This higher bit rate, almost error-free and low-delay communication network makes it possible to transmit high-definition content almost in real time [2]. In this environment, the use of lossless codecs (coders/decoders) is widespread these days [3]–[7]. Audio lossless coding can perfectly reconstruct the signal from the bit stream. Users are able to choose not only the efficiency of lossy coding, at the sacrifice of quality, but also the reliability of acoustic signals by lossless coding because the network bit rate and disk space are also increasing. However, since the size of bit streams encoded by a lossless coder depends on the character-

istics of the input signals, the lossless coding results in a variable bit rate that cannot be controlled. This is in contrast to lossy coding such as MPEG^{*1}-2/4 advanced audio coding (AAC) that is applied for portable music players and broadcasts [3], [4], [8]–[10].

One of the most promising applications of broadband networks including NGN is high-quality video and audio transmission. Digital cinema is one potential application of this, as well as content distribution from popular theaters and music halls. We sometimes refer to such a content delivery system as *other digital stuff* or *online digital sources* (ODS) [11], [12]. In Japan, several musical artists such as the Takarazuka

^{*1} The moving picture experts group (MPEG) is a working group of international organization for standardization (ISO) and the international electrotechnical commission (IEC) that develops standards for coded representation of digital audio and video and related data.

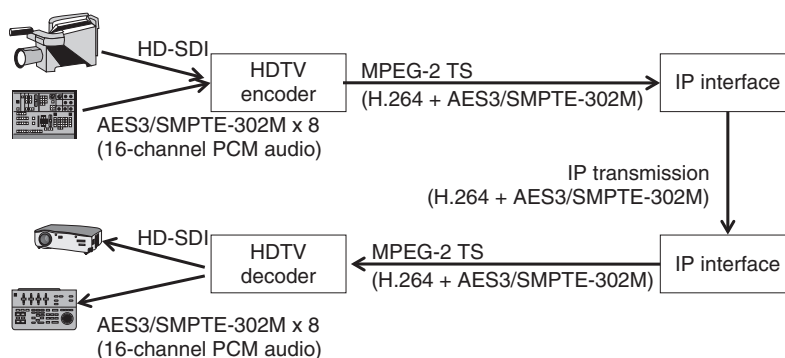


Fig. 1. System diagram of existing settings.

Revue Company [13], X-Japan, and L'Arc~en~Ciel [14] have provided live performances in real time to fans located in other places such as movie theaters. Video signals are normally compressed by ITU-T H.264/MPEG-4 AVC*², and audio signals by AAC to save the bit rate. Even though these transmissions are of music content, a lossy codec has been used for sound data. In addition, multichannel audio signals from the live stage are sometimes mixed down to two channels, and then the processed stereo data are transmitted to the other venue. Lossy compression and down-mixing are reasonable when the speaker settings are defined and the acoustic characteristics are the same in each place of delivery, but obviously such assumptions are not realistic. Consequently, there is room to improve the audio quality of ODS. One feasible idea is to transmit the sound of the musical instrument as-is (i.e., without down-mixing or lossy coding), and to down-mix at each local site.

To achieve a way to provide high-quality music, we carried out an experiment of lossless transmission of sound data. MPEG-4 Audio lossless coding (ALS) [15]–[17] was used to losslessly encode the 16-channel musical instrument data because it is an international standard technology and supports multichannel audio signals. The compressed audio data were transmitted from a live music venue (sender) to a café via an Internet protocol (IP) network. Although the bit rate of lossless audio became larger than that of the lossy one, we did not have to worry about degradation of sound waveforms. At the café, the transmitted sound data were decoded without any loss and appropriately remixed to adjust to the environment of the site. Lossless audio coding allowed the listeners in the offsite venue to enjoy the concert with customized audio data.

The experimental result from this trial motivated us to develop a new prototype codec that can control the bit rate of video and audio. Audio quality is guaranteed, and higher video quality is achieved by making use of extra bits saved by lossless audio compression.

The remainder of this article is as follows. We report on the lossless audio transmission experiment in section 2 and introduce the prototype codec in section 3. Finally, we conclude the article in section 4.

2. Investigation of lossless compression efficiency in demonstrative trial

2.1 Configuration

As described in the previous section, we carried out an experiment of lossless transmission of sound data to provide high-quality music. We used the high-definition television (HDTV) encoder/decoder (HV9100 series by NTT Electronics Corp. (NEL)) and the IP interface (NA5000 by NEL) as shown in **Figs. 1** and **2**. The HDTV encoder output an MPEG-2 transport stream (TS), which included eight pairs of AES3/SMPTE-302M*³ (i.e., 16-channel pulse-code modulation (PCM) audio signals) bitstreams that were input from a digital mixer and an H.264 bitstream that consisted of the encoded high-definition serial digital interface (HD-SDI) data from a high-vision camera.

*² ITU-T H.264/MPEG-4 AVC was prepared jointly by international telecommunication union telecommunication standardization sector (ITU-T) and by MPEG.

*³ AES3 is a standard that specifies the transport of digital audio signals between professional devices (published by the Audio Engineering Society); SMPTE-302M is a standard that specifies the transporting of AES3 data in an MPEG-2 transport stream for television applications (published by the Society of Motion Picture and Television Engineers).

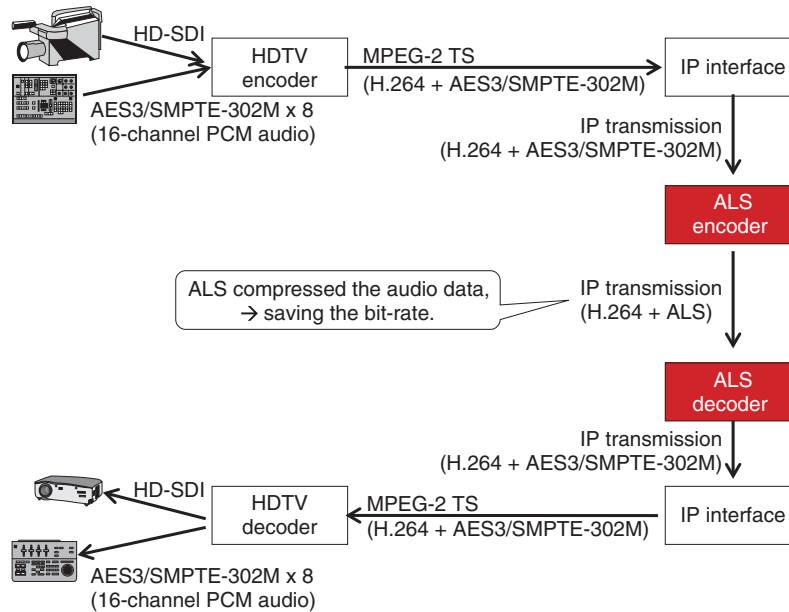


Fig. 2. System diagram of the experiment.

The MPEG-2 TS bitstream was converted to IP packets by the IP interface. Audio and video data were transmitted over the IP network. After transmission, an IP interface and HDTV decoder carried out the reverse operations. Finally, we obtained high-vision data and 16-channel PCM audio signals.

Although this existing setting (Fig. 1) can provide high-quality sound, it wastes network resources because there are redundant audio signals. From an ecological and economical viewpoint, it would be preferable to reduce the cost of the network without any degradation of audio quality.

For this experiment, we added the ALS encoder/decoder to save the bit rate, as shown in Fig. 2. Before IP transmission, the ALS encoder losslessly compressed eight pairs of the AES3/SMPTE-302M bitstreams. After the ALS decoder had received the encoded IP packets, it reconstructed the original IP packets and sent them to the IP interface. Then, the IP interface and HDTV decoder operated as if no processing had taken place.

The total bit rate was set to 50 Mbit/s (Fig. 3). The audio bit rate was about 18 Mbit/s because 16-channel signals were digitized at 48 kHz and 20 bits (four more bits were required for AES3/SMPTE-302M). Therefore, 32 Mbit/s was needed for the target bit rate of H.264 (video), which was not varied during the experiment. Then, the audio data were losslessly

encoded by ALS, and the bit rate of the ALS bitstream was reduced to less than 18 Mbit/s, depending on the input signals. The transmitted ALS bitstreams were perfectly decoded to PCM audio signals.

2.2 Experimental results

This experiment was conducted on March 19, 2009 from 7:00 to 10:00 PM. There were about 80 listeners at the live venue and around 100 at the café. The sender and receiver sites are shown in Figs. 4 and 5, respectively.

The bit rate of the ALS bitstreams on average for each minute is shown in Fig. 6. The audio signals are losslessly compressed to 21% (3.78 Mbit/s)–60% (10.87 Mbit/s) of the original data size. The bit rate increased as the performance progressed because the number of musicians (i.e., musical instruments) also increased. In summary, the ALS was able to save about 11.2 Mbit/s on average, so the total bit rate became around 40 Mbit/s.

Sound data were appropriately remixed for adjustment to the environment of the receiving site. In this experiment, it was easy for the mixing engineers (also known as public address or sound reinforcement engineers) at the café to carry out their remixing tasks because the original sound data came from the live venue.

We received positive feedback from the invited

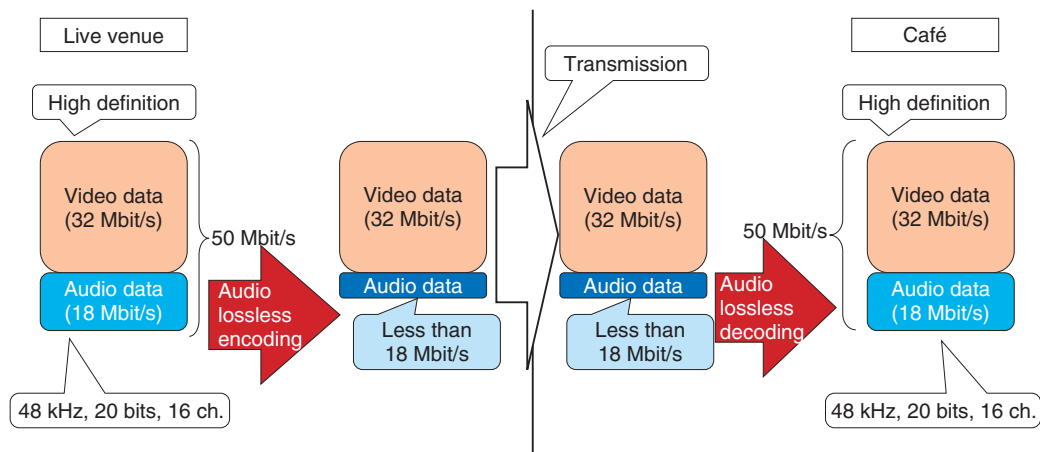


Fig. 3. Conceptual diagram of the experiment.



Fig. 4. Live venue in Shimokitazawa, Tokyo (sender site).



Fig. 5. Café in Aoyama, Tokyo (receiver site).

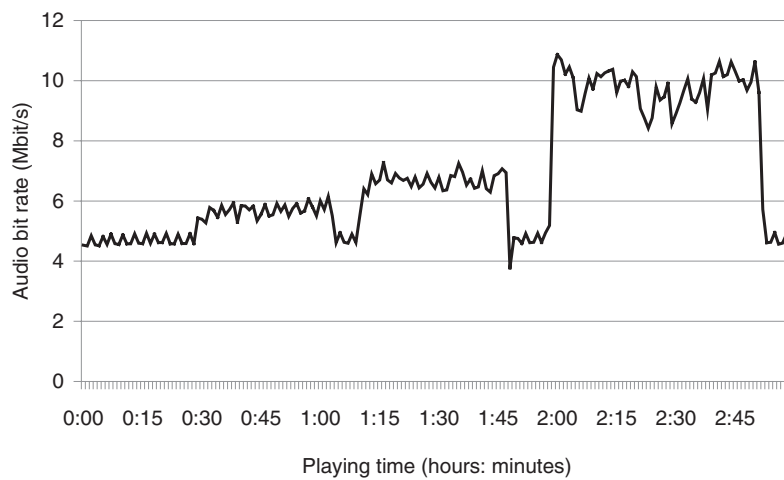


Fig. 6. Bitrate of compressed audio.

guests, and we believe that assigning a higher bit rate for multichannel audio signals was key in the success of the experiment. Our experiment suggests that the combination of high-definition video and audio data, especially ODS with lossless audio coding such as the ALS, will enable fans to enjoy live musical performances from offsite locations. We think this trial is a good example of a content delivery service with high-quality audio. As far as we know, this is the first experiment on real-time lossless audio transmission (especially ALS) with HD video via an IP network.

3. Prototype of audio rate-oriented adaptive bit-rate video codec

3.1 Concept of the developed codec

The results of the experiment described in the previous section indicate that we can reduce the bit rate of audio signals by approximately half by using ALS. For real-time streaming, we should nevertheless retain the worst-case bit rate (i.e., PCM rate) for audio. We developed a prototype of an audio bit-rate-oriented adaptive bit-rate video codec to determine whether we could make efficient use of the reduced bit rate, as shown in **Fig. 7**. The bit rate saved by using ALS can then be contributed to H.264.

Let V be the bit rate of video, A be that of audio, and C be that of losslessly compressed audio (i.e., $C \leq A$). As shown in **Fig. 8**, the conventional codec needs $V + A$ bit/s to transmit the data. By contrast, the developed codec can provide the additional bits for the video codec. Therefore, the video codec can use $V + (A - C)$ bit/s, and the audio codec needs only C bit/s. The video quality is therefore enhanced. The total bit rate for the TS is the same, $V + A$ bit/s, and the audio quality is also the same because we use lossless coding.

3.2 Preliminary experiment with the developed codec

The prototype encoder supports.

- HD-SDI signal input → H.264 bitstream output
- Three AES3/SMPTE-302M inputs (i.e., up to 6 channels) → MPEG-4 ALS bitstream output.

The decoder supports.

- H.264 bitstream input → HD-SDI signal output
- MPEG-4 ALS bitstream input → three AES3/SMPTE-302M outputs.

In the preliminary experiment with the developed codec, the TS rate was set to 25 Mbit/s because we assumed 30 Mbit/s for the NGN operation for IP transmission (which is reasonable for users) and because forward error correction usually requires

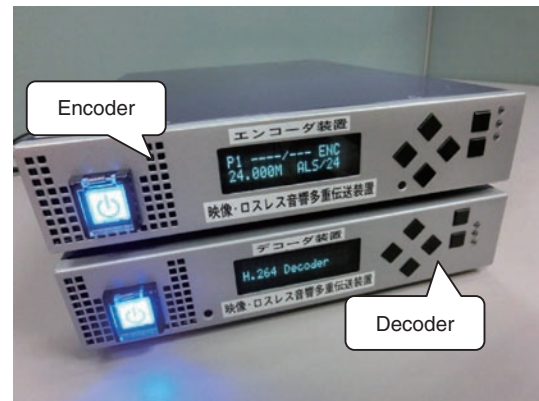


Fig. 7. The fabricated encoder and decoder.

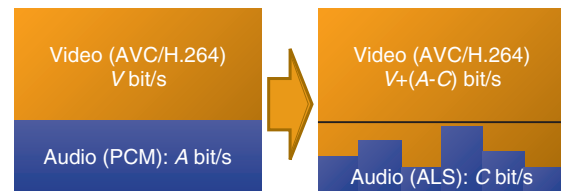


Fig. 8. Conceptual diagram of prototype codec.

10% of the bit rate. Content from an opera performed and provided by a famous Japanese opera company was used as input signals.

A diagram of the settings for the preliminary experiment is shown in **Fig. 9**. A Blu-ray player output an HDMI (high-definition multimedia interface) signal, and the converter divided it into HD-SDI and AES3/SMPTE-302M signals. The prototype encoder produced the H.264 bitstream from the obtained HD-SDI signal, of which the target bit rate depends on the bit rate of audio data compressed by the ALS, and produced the ALS bitstream from the AES3/SMPTE-302M signals. The decoder then output the HD-SDI signal and reconstructed the AES3/SMPTE-302M signals losslessly.

We captured TS bitstreams and analyzed them in order to evaluate the performance of the fabricated codec. Without lossless audio coding, the bit rate of audio requires 6.9 Mbit/s (48 kHz x 24 bits x 6 channels), which means that the remaining bit rate is only 18 Mbit/s for video. In contrast, as shown in **Fig. 10**, the bit rate of video can use more than 18 Mbit/s (statistically 18.40 (mean) \pm 0.55 (standard deviation) Mbit/s) because the bit rate of audio is

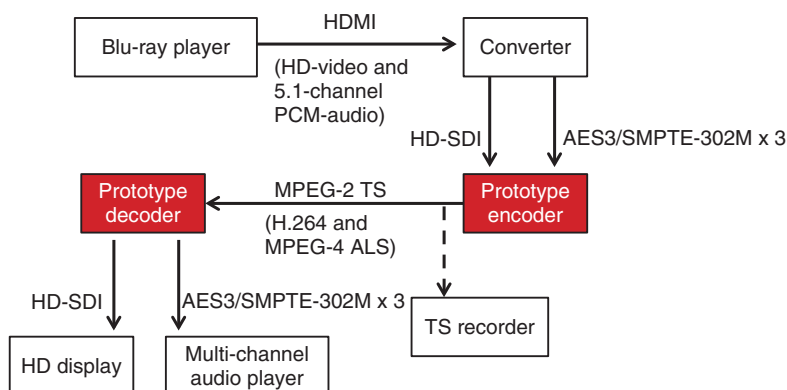


Fig. 9. Configuration for preliminary experiment.

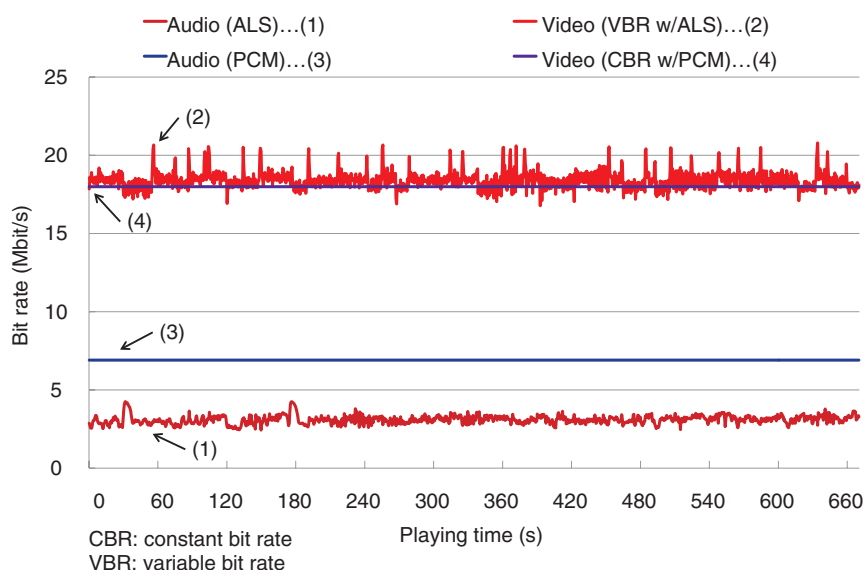


Fig. 10. Bit rates in preliminary experiment.

compressed to around 3 Mbit/s (statistically 3.10 (mean) \pm 0.26 (standard deviation) Mbit/s). In summary, video quality is improved by utilizing lossless audio compression.

This codec is still a prototype, so we decided to use low-risk, low-return settings for the bit rate and adaptive bit-rate control. We can achieve better quality by fine-tuning these settings.

4. Conclusion

We conducted an experiment of lossless audio transmission via an IP network. Multichannel audio

signals compressed by MPEG-4 ALS and high-definition video signals were transmitted from a live venue to a café to provide high-quality music for an off-site audience. This experimental result motivated us to develop a prototype codec that controls the bit rate between video and audio. Audio quality is guaranteed, and higher video quality is achieved by making use of extra bits saved by the lossless audio compression, while the video encoder is able to use the remaining bits subject to the constant total bit rates. After the prototype codec is refined, we will be able to transmit higher quality content such as operas, musicals, and concerts. Japanese broadcasting

standards support MPEG-4 ALS for downloading 22.2-channel high-definition audio [18], and the developed audio/video codec was used to transmit the high-quality multichannel audio signal of a Takarazuka Revue performance [19]. Thus, the described technologies are expected to be widely used in the near future.

Acknowledgements

We thank all of the engineers who assisted with the development and experiments.

References

- [1] S. Esaki, A. Kurokawa, and K. Matsumoto, "Overview of the Next Generation Network," NTT Technical Review, Vol. 5, No. 6, 2007. <https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200706sf1.html>
- [2] K. Kawazoe, R. Kakinuma, Y. Haneda, D. Minoura, S. Minamoto, and H. Ishimoto, "Platform Application Technology Using the Next Generation Network," NTT Technical Review, Vol. 5, No. 6, 2007. <https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200706sf3.html>
- [3] A. Spanias, T. Painter, and V. Atti, "Audio Signal Processing and Coding," John Wiley & Sons, Inc., 2007.
- [4] S. Salomon and G. Motta, "Handbook of Data Compression," Springer, 2010.
- [5] M. Hans and R. W. Schafer, "Lossless Compression of Digital Audio," IEEE Signal Processing Magazine, Vol. 18, No. 4, pp. 21–32, 2001.
- [6] K. Konstantinides, "An Introduction to Super Audio CD and DVD-Audio," IEEE Signal Processing Magazine, Vol. 20, No. 4, pp. 71–82, 2003.
- [7] B. H. Kuzuki, N. Fuchigami, and J. R. Stuart, "DVD-Audio Specifications," IEEE Signal Processing Magazine, Vol. 20, No. 4, pp. 72–90, 2003.
- [8] ISO/IEC 13818-7, "Information technology—Generic coding of moving pictures and associated audio information—Part 7: Advanced Audio Coding (AAC)," Dec. 1997.
- [9] ISO/IEC 14496-3, "Information technology—Coding of audio-visual objects—Part 3: Audio," Dec. 1999.
- [10] ARIB STD-B32, "Video Coding, Audio Coding and Multiplexing Specifications for Digital Broadcasting," May 2001.
- [11] S. Kim, M. Ogawara, T. Fujii, Y. Kamamoto, N. Harada, and T. Moriya, "Requirements for Developing Ultra-Realistic Live Streaming Systems," Proc. of the 2009 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2009), pp. 175–178, Kanazawa, Japan.
- [12] H. Takahashi, D. Shirai, T. Murooka, and T. Fujii, "Multipoint Streaming Technology for 4K Super-high-definition Motion Pictures," NTT Technical Review, Vol. 5, No. 5, 2007. <https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200705le1.html>
- [13] Press release on live video transmission of Takarazuka Review performance (in Japanese). <http://www.ntt.co.jp/news/news07/0711/071114a.html>
- [14] Information on live video transmissions of L'Arc-en-ciel performance (in Japanese). http://www.larc-en-ciel.com/m/news/larc/other/l_ot_ar-08.html
- [15] Information technology—Coding of audio-visual objects—Part 3 Audio, 3rd Ed. Amendment 2: Audio Lossless Coding (ALS), new audio profiles and BSAC extensions, ISO/IEC Std. 14496-3:2005/AMD.2:2006, March 2006.
- [16] T. Liebchen, T. Moriya, N. Harada, Y. Kamamoto, and Y. Reznik, "The MPEG-4 Audio Lossless Coding (ALS) standard—Technology and applications," Proc. of the 119th Audio Engineering Society Convention, Paper #6589, New York, NY, USA, 2005.
- [17] Y. Kamamoto, N. Harada, T. Moriya, S. Kim, M. Ogawara, and T. Fujii, "Experiment of Sixteen-Channel Audio Transmission Over IP Network by MPEG-4 ALS and Audio Rate-Oriented Adaptive Bit-Rate Video Codec," Proc. of the 129th Audio Engineering Society Convention, Paper #8302, San Francisco, CA, USA, 2010.
- [18] ARIB STD-B45, "Content Download System for Digital Broadcasting," Apr. 2010.
- [19] H. Yamane, A. Yamashita, K. Kamatani, M. Morisaki, T. Mitsunari, and A. Omoto, "High-presence Audio Live Distribution Trial," NTT Technical Review, Vol. 9, No. 10, 2011. <https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201110fa4.html>



Yutaka Kamamoto

Research Scientist, Moriya Research Laboratory, NTT Communication Science Laboratories.

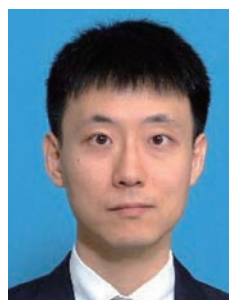
He received the B.S. degree in applied physics and physico-informatics from Keio University, Kanagawa, in 2003 and the M.S. and Ph.D. degrees in information physics and computing from the University of Tokyo in 2005 and 2012, respectively. Since joining NTT Communication Science Laboratories in 2005, he has been studying signal processing and information theory, particularly lossless coding of time-domain signals. He additionally joined NTT Network Innovation Laboratories, where he developed the audio-visual codec for ODS from 2009 to 2011. He has contributed to the standardization of coding schemes for MPEG-4 Audio lossless coding (ALS) and ITU-T Recommendation G.711.0 Lossless compression of G.711 pulse code modulation. He received the Telecom System Student Award from the Telecommunications Advancement Foundation (TAF) in 2006, the IPSJ Best Paper Award from the Information Processing Society of Japan (IPSJ) in 2006, the Telecom System Encouragement Award from TAF in 2007, and the Awaya Prize Young Researcher's Award from the Acoustical Society of Japan (ASJ) in 2011. He is a member of IPSJ, ASJ, the Institute of Electronics, Information and Communication Engineers (IEICE), and IEEE.



Sunyong Kim

Research Engineer, Media Innovation Laboratory, NTT Network Innovation Laboratories.

She received the B.E. and M.E. degrees in information science from the University of Tokyo in 2002 and 2004, respectively. She joined NTT in 2004 and has been researching conversation scene analysis on highly realistic remote collaboration systems. She is a member of IPSJ and IEICE.



Takahiro Yamaguchi

Senior Research Engineer, Media Innovation Laboratory, NTT Network Innovation Laboratories.

He received the B.E., M.E., and Ph.D. degrees in electronic engineering from the University of Electro-Communications, Tokyo, in 1991, 1993, and 1998, respectively. He joined NTT in 1998 and has been researching super-high-definition image distribution systems. He is a member of IEICE and the Institute of Image Information and Television Engineers (ITE).



Noboru Harada

Senior Research Scientist, Moriya Research Laboratory, NTT Communication Science Laboratories.

He received the B.S. and M.S. degrees from the Department of Computer Science and Systems Engineering of Kyushu Institute of Technology, in 1995 and 1997, respectively. He joined NTT in 1997. His main research area has been lossless audio coding, high-efficiency coding of speech and audio, and their applications. He additionally joined NTT Network Innovation Laboratories, where he developed the audio-visual codec for ODS, from 2009 to 2011. He is an editor of ISO/IEC 23000-6:2009 Professional Archival Application Format, ISO/IEC 14496-5:2001/Amd.10:2007 reference software MPEG-4 ALS, and ITU-T G.711.0. He is a member of IEICE, ASJ, the Audio Engineering Society (AES), and IEEE.



Masanori Ogawara

Senior Research Engineer, Supervisor, Media Innovation Laboratory, NTT Network Innovation Laboratories.

He received the B.E. and M.E. degrees in electrical engineering from Keio University, Kanagawa, in 1992 and 1994, respectively. He joined NTT in 1994. His current research interests include reliable IP transmission technologies and network supported content creation collaboration systems. He is the director of a collaborative working platform development project. He received a paper award from IEICE in 1999. He is a member of IEICE.



Takehiro Moriya

NTT Fellow, Moriya Research Laboratory, NTT Communication Science Laboratories.

He received the B.S., M.S., and Ph.D. degrees in mathematical engineering and instrumentation physics from the University of Tokyo in 1978, 1980, and 1989, respectively. Since joining the Nippon Telegraph and Telephone Public Corporation (now NTT) in 1980, he has been engaged in research on medium to low bitrate speech and audio coding. In 1989, he worked at AT&T Bell Laboratories, NJ, USA, as a visiting researcher. Since 1990, he has contributed to the standardization of coding schemes for the Japanese Public Digital Cellular system, ITU-T G.729 and G.711.0, and ISO/IEC MPEG MPEG-4 General Audio and MPEG-4 ALS. He is a Fellow member of IEEE and a member of IPSJ, IEICE, AES, and ASJ.



Tatsuya Fujii

Senior Research Engineer, Supervisor, Group Leader of Media Processing Systems Research Group, Media Innovation Laboratory, NTT Network Innovation Laboratories.

He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Tokyo in 1986, 1988, and 1991, respectively. He joined NTT in 1991. He has been researching parallel image processing and super-high-definition image communication networks. In 1996, he was a visiting researcher at Washington University in St. Louis, MO, USA. He is a member of IEICE, ITE, and IEEE.

Distributed Array Antenna Technique for Satellite Communications

*Kouhei Suzuki, Yoshinori Suzuki,
Takashi Hirose, and Takatoshi Sugiyama*

Abstract

NTT Access Network Service Systems Laboratories has developed a distributed array antenna (DAA) technique to achieve a higher gain earth station antenna for satellite communications. The DAA configuration comprises several smaller antennas and has an antenna gain equivalent to that of a conventional large-aperture antenna. We applied the configuration to two antennas and experimentally confirmed that the DAA achieves increased antenna gain.

1. Introduction

Satellite communication systems are utilized to provide temporary communication lines when a disaster occurs. An example is the system established by the NTT Group that uses the Ku-band (14 GHz/12 GHz) for communications to provide relief in the early stages of post-disaster recovery.

Many different types of earth stations are used for disaster recovery, such as those shown in **Fig. 1** [1], [2]. Terminal earth stations are deployed at evacuation centers, and the base station aggregates the communication lines from these terminal earth stations, as shown in **Fig. 2**.

In stricken areas in particular, the earth stations need to make use of small-aperture antennas that can be transported easily. One problem with these stations is that the maximum number of available telephone lines and the maximum transmission bit rate are limited because the antenna gain is proportional to the antenna aperture. However, large-aperture antennas are heavy to transport and must be installed by a skilled engineer.





A further problem is that the base station has a very large and heavy aperture antenna in order to obtain high antenna gain. A certain amount of space is required on building rooftops for this kind of antenna, and such space is also limited. Therefore, the base station antenna should consist of relatively small and light antennas that can be easily placed.

2. Proposed configuration

To address these problems, NTT Access Network Service Systems Laboratories proposes a distributed array antenna (DAA) technique and configuration, shown in **Fig. 3**. The DAA configuration, which comprises several smaller antennas, has an antenna gain equivalent to that of a conventional larger aperture antenna. If we apply this DAA configuration to three 0.75-m-diameter antennas (**Fig. 1**), we can achieve gain almost equal to that of a 1.2-m-diameter antenna. Furthermore, the base station can achieve the antenna gain of a 4.5-m-diameter antenna that has a configuration of four 2.4-m-diameter antennas. The total weight of these antennas is 1200 kg, which makes it smaller and lighter than a conventional base station antenna. Thus, there is greater flexibility in positioning it because it requires less installation space. Another benefit is that the use of multiple antennas improves failure resistance; if one antenna fails, the others can maintain the communication even with a lower bitrate.

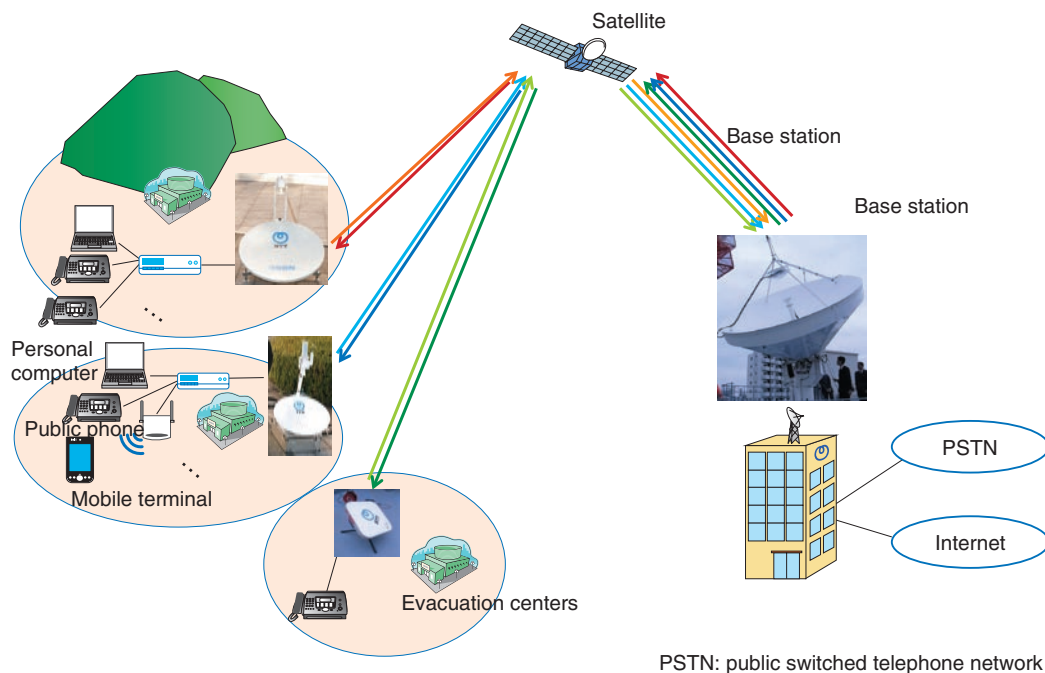
3. Key technique

The DAA system consists of a modem, several radio frequency (RF) units, and a DAA controller, as shown in **Fig. 4**. Each RF unit includes an antenna, a frequency duplexer, a block upconverter (BUC), and a low noise block converter (LNB). The BUC converts

	Terminal earth stations			Base station
	Ku-band ultrasmall earth station	Small satellite station developed in 2012	Portable earth station	
Antenna				
Aperture	0.55 m	0.75 m	1.2 m	4.5 m
Weight	30 kg	40 kg	150 kg	Over 2000 kg
Number of available telephone lines	1	10	40	Over 40
Transmission capacity	35.5 kbit/s (16-kbit/s LD-CELP)	384 kbit/s	1.5 Mbit/s	Over 6 Mbit/s

LD-CELP: low-delay code excited linear prediction

Fig. 1. Existing satellite earth stations used by NTT.



PSTN: public switched telephone network

Fig. 2. Existing satellite communication system image.

the carrier frequency from the modem (1 GHz) to the transmission frequency band (14 GHz) and amplifies the transmitted signal level for the uplink. The satellite relays these signals while converting the frequen-

cy from the 14-GHz band to the 12-GHz band, and the LNB amplifies the received signals and converts the frequency from the received frequency band (12 GHz) to the modem (1 GHz).

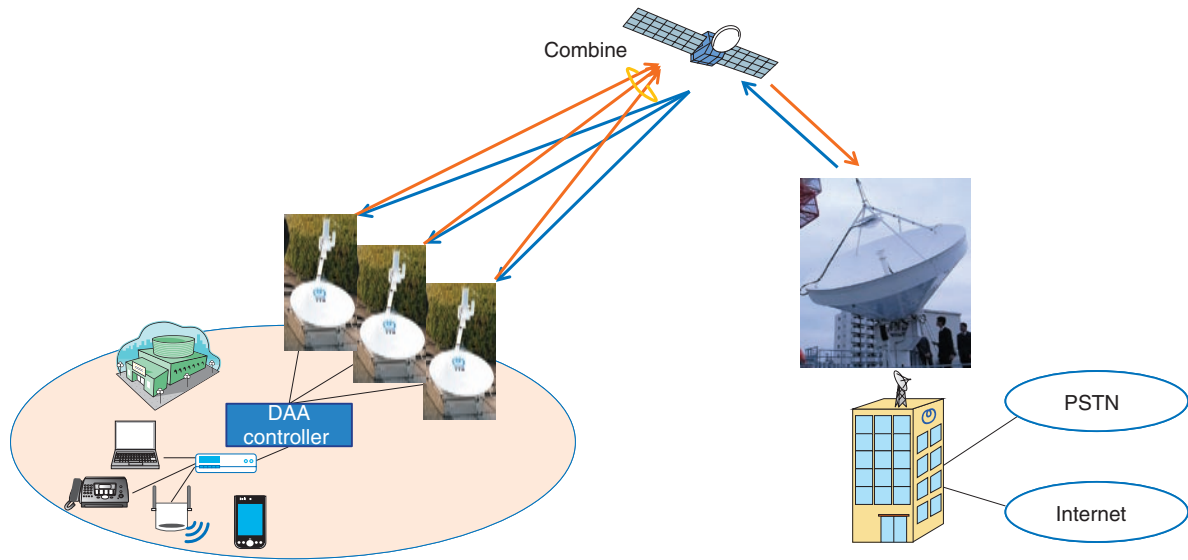


Fig. 3. DAA system image of terminal earth station.

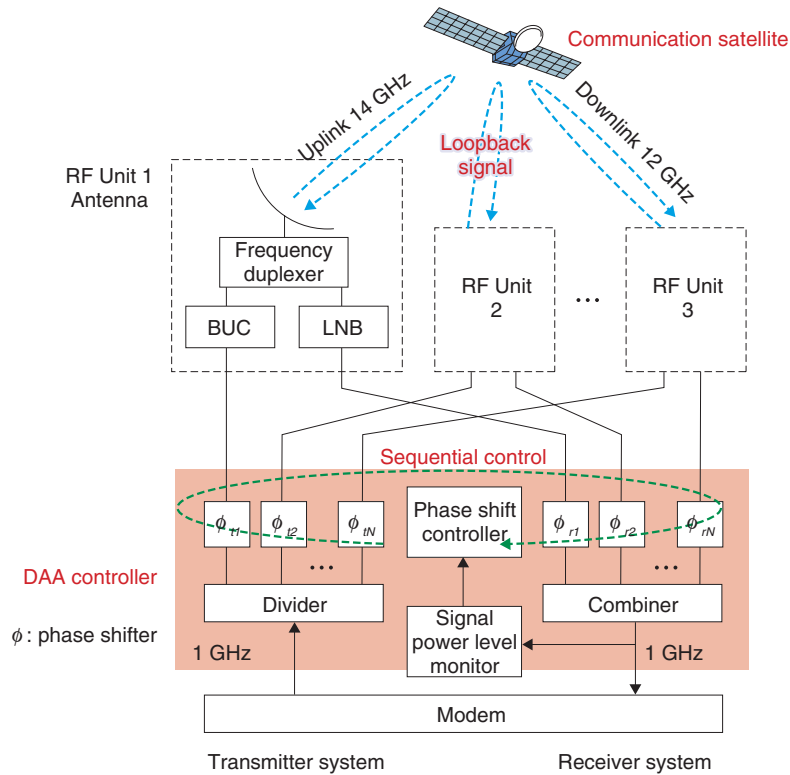


Fig. 4. DAA system configuration.

To obtain a higher antenna gain, it is essential to maintain in-phase combining at the satellite onboard antenna in the DAA uplink. The modem output signal is first divided by a divider. Then, the phase of the signal is calibrated by a phase shifter, and the signal is supplied to the antennas. If a phase error occurs between the signals transmitted by the RF units, it will degrade the composite antenna gain, as shown in **Fig. 5**.

There are two possible causes of phase errors in the DAA uplink: the location of the RF unit and slow fluctuation of BUC characteristics over time. Generally, the conventional configuration has a feedback circuit to compensate for these phase errors [3]. However, the proposed DAA does not require a feedback circuit. The DAA compensates for this by having stations transmit their own signals to the satellite and then receive them back from the satellite. During that time, the phase shifters calibrate the phase to maintain the loopback signal at a maximum level.

The two kinds of phase errors can be calibrated comprehensively because the phase error caused by the RF unit location consists of static phase offset, and the phase error caused by BUC fluctuation is slower than the satellite delay (250 ms).

Specifically, the phase calibration is carried out by selecting phase shifters one-by-one. The selected phase shifter shifts phases at a constant positive or negative value. Simultaneously, the loopback signal level is registered by a monitor. Then, the DAA controller chooses the set phase that records the highest signal level.

The features of the proposed system are:

- No feedback signal generator for transmitted signals is required.
- There is no need to remodel the RF units or modem because a controller is inserted between a general purpose modem and the RF units.
- The two abovementioned causes of phase errors can be compensated as a whole.

In the downlink, the DAA controller combines the signals from each RF unit so as to maximize the combined received signal level. There are also two possible causes of phase errors in the downlink: RF unit location and slow fluctuation of LNB characteristics with time. These phase errors can be compensated for in a similar manner.

4. Satellite experiments

We configured the DAA using two commercially available RF units in 1.2-m-diameter antennas, shown

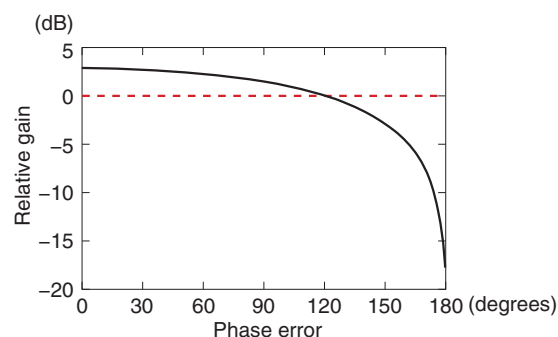


Fig. 5. Phase error vs. relative gain.



Fig. 6. Appearance of DAA.

in **Fig. 6**, and we used the Ku-band satellite service to evaluate the phase calibration method, antenna gain, and transmission characteristics of the signal constellation and the bit error rate (BER) performance.

We conducted a test to evaluate the phase calibration method. The loopback signal level results with and without calibration are shown in **Fig. 7**. For the test without calibration, the calibration was turned off at elapsed time 0. We can confirm from the results that the received signal level is degraded. By contrast, with our calibration method the signal level is kept constant.

With the DAA, the gain is twice as high as that of the 1.2-m-diameter antenna, and the BUC output power is also twice as high, theoretically a fourfold (6 dB) improvement. To evaluate the DAA gain, we compared the received signal spectra of the DAA, a 1.2-m-diameter antenna, and a 1.8-m-diameter antenna (**Fig. 8**). The 1.8-m-diameter antenna had twice

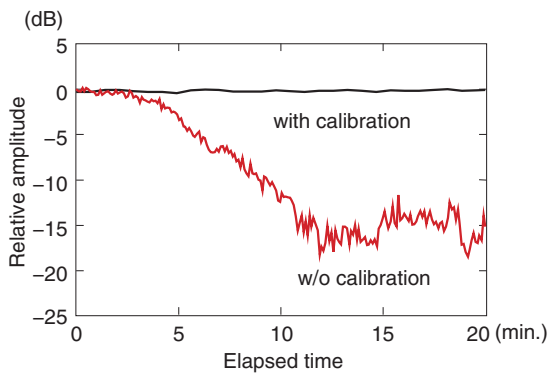


Fig. 7. Received signal levels with and without calibration.

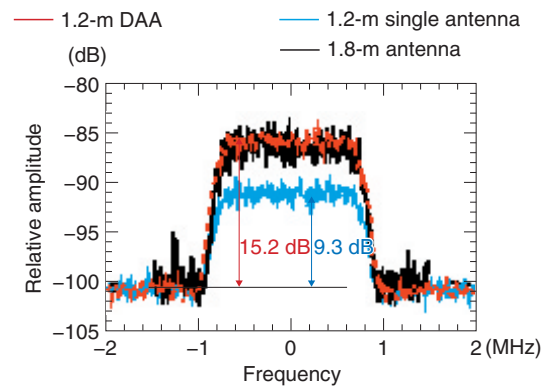


Fig. 8. Received signal spectra.

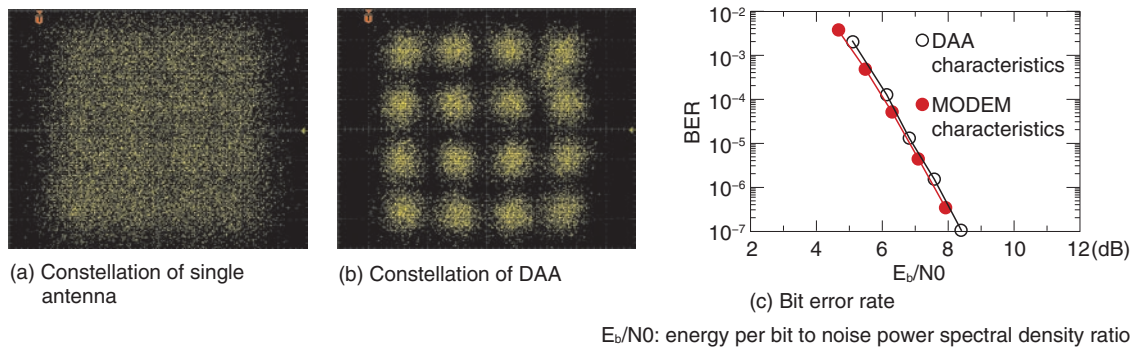


Fig. 9. Signal transmission characteristics.

the gain of the 1.2-m-diameter antenna. To establish equal conditions, we used a 1.8-m-diameter antenna that had a BUC whose output power was twice as high as that of the BUC in the smaller antenna. The results show that the received signal spectrum of the DAA is almost the same as that of the 1.8-m-diameter antenna and is 5.9 dB higher than that of the 1.2-m-diameter antenna. These results indicate that the degradation from the theoretical value is only 0.1 dB.

Finally, we evaluated the transmission characteristics of 3.2-Mbit/s, 16-quadrature amplitude modulation (QAM) signals. The 16-QAM scheme has a 4x4 signal constellation, and if there is sufficient transmission energy, a constellation can be easily distinguished. The signal constellations for the 1.2-m-diameter antenna and the DAA are shown in **Figs. 9(a)** and **(b)**, respectively. A 4x4 signal constellation was impossible to distinguish for the 1.2-m-diameter antenna because there was insufficient transmission energy. However, because of the increased gain in the

DAA, the transmitted power was high enough to enable a 4x4 signal constellation to be distinguished. The measured BER performance is shown in **Fig. 9(c)**. The resulting BER performance is close to that of the case using a direct cable connection (MODEM).

5. Conclusions

We have developed a DAA technique that comprises several small-aperture antennas and that can achieve gain equivalent to that of a large-aperture antenna. The DAA controller can be easily introduced in existing satellite earth station antennas and modems to increase the antenna gain. Experimental results for Ku-band satellite communication confirmed that the DAA increased the gain. In the future we plan to enhance our system in order to extend its usability to areas such as maritime mobile communications.

References

- [1] Y. Imaizumi, T. Hirose, and H. Yoshida, "Small Satellite Earth Stations for Disaster Recover Operations," NTT Technical Review, Vol. 10, No. 7, 2012.
<https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201207ra2.html>
- [2] T. Manabe, "Trends in Wireless Access Technologies toward Expansion of Broadband and Ubiquitous Services," NTT Technical Review, Vol. 9, No. 5, 2011.
<https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201105fa4.html>
- [3] K. Suzuki, Y. Suzuki, and K. Kobayashi, "A Novel Earth Station Antenna Concept for Ku-band Mobile Satellite Communication Systems—Distributed Array Antenna and Key Technologies," Proc. of the 29th AIAA International Communications Satellite Systems Conference (ICSSC-2011), Nara, Japan.



Kouhei Suzuki

Researcher, Satellite Communication Systems Group, Wireless Access Systems Project, NTT Access Network Service Systems Laboratories.

He received the B.E. and M.E. degrees in electrical engineering from Tokyo University of Science in 2007 and 2009, respectively. Since joining NTT Access Network Service Systems Laboratories in 2009, he has been engaged in research on signal processing of the Satellite Earth Station.



Takashi Hirose

Senior Research Engineer, Supervisor, Satellite Communication Systems Group, Wireless Access Systems Project, NTT Access Network Service Systems Laboratories.

He received the B.E. and M.E. degrees in mechanical engineering from Keio University, Kanagawa, in 1989 and 1991, respectively. Since joining NTT Telecommunication Networks Laboratories in 1991, he has been engaged in the development of a routing system for an asymmetrical satellite Internet communication system. He is currently working on next-generation satellite communication systems for disaster recovery.



Yoshinori Suzuki

Senior Research Engineer, Satellite Communication Systems Group, Wireless Access Systems Project, NTT Access Network Service Systems Laboratories.

He received the B.E., M.E., and Ph.D. degrees from Tohoku University, Miyagi, in 1993, 1995, and 2005, respectively. Since joining NTT Wireless Systems Laboratories in 1995, he has been engaged in R&D of multi-beam antenna feed systems for communication satellites. He received the Young Researcher's Award from IEICE in 2002 and the Best Paper Award of the NTT Technical Publications in 2007. He is a member of IEICE.



Takatoshi Sugiyama

Senior Research Engineer, Supervisor, Group Leader of Satellite Communication Systems Group, NTT Access Network Service Systems Laboratories.

Since joining NTT in 1989, he has been engaged in R&D of forward error correction, interference compensation, CDMA, modulation-demodulation, and MIMO-OFDM technologies for wireless communication systems such as satellites, wireless ATM, wireless LAN, and cellular systems.

Development of ITU-T Action Plans for New Study Period at WTSA-12

Hideo Imanaka and Yoshinori Goto

Abstract

WTSA, the World Telecommunication Standardization Assembly, is the highest-level assembly of the International Telecommunication Union, Telecommunication Standardization Sector (ITU-T). The most recent assembly, WTSA-12, was held in Dubai, United Arab Emirates (UAE), in November 2012. The purpose of the assembly was to develop the ITU-T action plans for the new study period from 2013 to 2016. Three Study Group vice-chairmen were elected from NTT. This article introduces the major results of WTSA-12.

1. Overview of ITU-T WTSA-12

1.1 Overview of WTSA

The organizational structure of the International Telecommunication Union (ITU), is illustrated in Fig. 1. Just as the three sectors of ITU, ITU-T (Telecommunication Standardization Sector), ITU-R (Radiocommunication Sector), and ITU-D (Development Sector), are placed under the General Secretariat, the high level conferences and assemblies are organized under the Plenipotentiary Conference, which is the highest-level meeting for decision making.

WTSA (World Telecommunication Standardization Assembly) is the assembly dealing with ITU-T matters. Similarly, WTDC (World Telecommunication Development Conference) and RA (Radiocommunication Assembly) are associated with ITU-D and ITU-R, respectively. In addition, WRC (World Radiocommunication Conference) is held to discuss the international assignment of radio frequencies, and WCIT (World Conference on International Telecommunication) is held to discuss the International Telecommunication Regulations (ITRs).

WTSA is held every four years to discuss issues concerning Study Group (SG) structure, the appointment of SG chairmen and vice-chairmen, the approval of Recommendations requiring a decision at a higher level than the SG, and the development and revision of Resolutions. Because of the importance of

the issues to be dealt with, regional preparatory meetings, which typically get started two years before the WTSA, are organized in order to coordinate the regional opinions of all six member regions. Japan belongs to the Asia-Pacific region, so delegates from Japan attend the Asia-Pacific regional preparatory meetings organized by the Asia-Pacific Telecommunity (APT). APT was established to support the development of telecommunication industries in the Asia-Pacific region through the participation of 38 countries including Japan, as shown on the right side of Fig. 1.

The WTSA-12 preparatory meeting in APT consisted of four Working Groups (WGs); one of the groups discussed SG structure and was led by one of the authors, Mr. Goto, as a rapporteur. As shown on the right side of Fig. 1, the APT common proposals were made from the results of preparatory meeting discussions. These common proposals were submitted to WTSA-12 representing the common opinions from the 38 countries in APT. There were 14 APT common proposals presented at WTSA-12.

1.2 Structure of WTSA-12

Around 700 people from 102 countries participated in WTSA-12 [1] held in Dubai, United Arab Emirates (UAE), from November 20–30, 2012. There were 32 Japanese delegates, including 6 from the NTT group.

Because of the wide range of discussion topics, five committees were organized to address specific topics

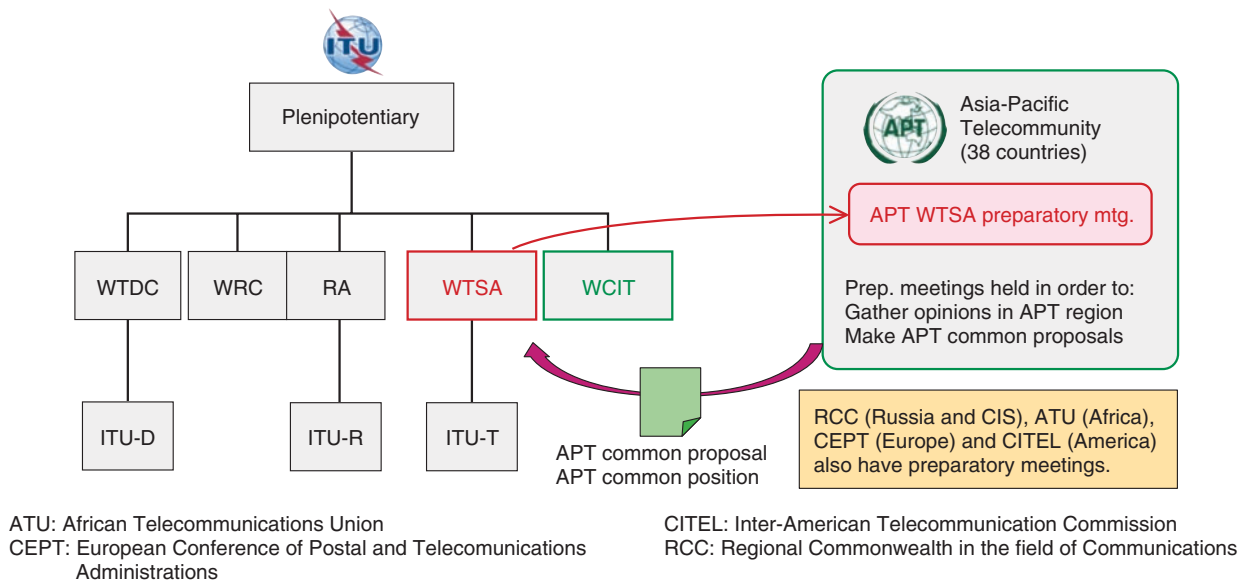


Fig. 1. Positioning of WTSA and preparatory meeting within APT.

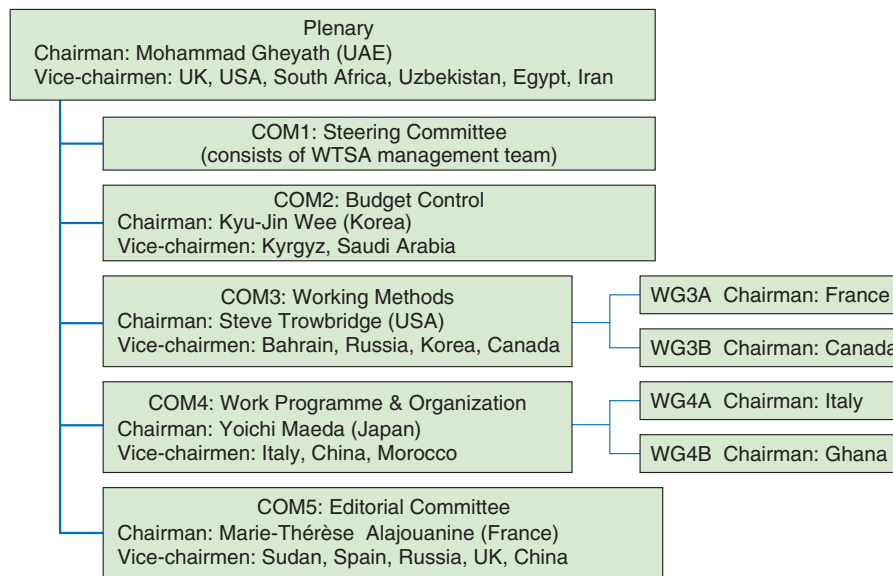


Fig. 2. Structure of WTSA-12.

under a Plenary that was led by the UAE, as shown in **Fig. 2**. Committee 4, referred to as Com4, was responsible for discussing SG structure and important issues for ITU-T organizations of standardization work, and was chaired by Mr. Yoichi Maeda of the Telecommunications Technology Committee (TTC) of Japan. Since Com4 dealt with more than half of all contributions submitted to WTSA-12, two WGs were created under Com4 to discuss issues separately. In

addition, several ad-hoc groups were set up that discussed inconclusive issues of WGs. Active discussions were continued from early morning to midnight throughout the weekend of the conference.

At this WTSA, discussions were also held on issues concerning numbering associated with the revision of ITRs that would be discussed at WCIT held just after WTSA.

2. SG structure and chairmen and vice-chairmen

2.1 SG structure

At the previous WTSA in 2008, the merger of SGs was discussed. The current structure of 10 SGs resulting from WTSA-08 was the outcome of transferring the mandates of SG6 and 19 to SG15 and 13, respectively, while splitting the mandates of SG4 between SG2 and 15. At WTSA-12, the APT proposed maintaining the current 10 SGs in recognition of the leading role of ITU-T in the global standardization of information and communications technology (ICT). This proposal gained support from other regions and countries including Africa, the Arab region, and Russia, while the European region showed their preference for reducing the number of SGs. It was agreed to keep the current 10 SGs.

The European region also proposed merging, deleting, and/or transferring low activity Questions. Delegates familiar with the situations of the Questions provided some explanations and justification for maintaining them. It was agreed to keep the structure as proposed by the SGs, and it was also agreed to request the SG in question to report the measures they would take to enhance their activities in order to justify the continuation of such Questions at the TSAG (Telecommunications Standardization Advisory Group) meeting in 2014.

2.2 Mandate of SGs

Several proposals regarding study areas in the next study period were made from delegates from APT and other regions. These proposals were considered

for the revision of Resolution 2, which describes new mandates for each SG. WTSA also adopted a new resolution on Software Defined Networking (SDN) that was proposed by APT. The aim of this new resolution is to encourage the study of SDN by ITU-T and to highlight the work of SG13 (Future Networks).

The APT common proposal, led by Japan, proposed assigning new study areas such as disaster-related issues and interoperability testing to the appropriate SGs and was basically agreed. The issue on sharing security work in cloud computing was discussed intensively, primarily between the USA and Russia involving the chairmen of the SGs in question. It was confirmed that the current responsibilities would be maintained; that is, SG17 is the lead SG on security issues, and SG13 is the lead SG on cloud computing. However, the discussion on the responsibilities concerning the security aspects of cloud computing was to be continued between the relevant parties until the TSAG meeting in 2013.

These results were reflected in the revised Resolution 2 and adopted at the plenary meeting.

2.3 Chairmen and vice-chairmen of SGs

Resolution 35, which specifies the terms of chairmen and vice-chairmen of SGs, was also adopted in this WTSA. At the Head of Delegation meeting, which was an unofficial meeting called by the ITU-T Secretariat Bureau and held during the latter half of WTSA-12, 16 chairmen including 10 SG chairmen and 6 regional tariff group chairmen, and 107 vice-chairmen were elected. From Japan, 3 chairmen and 8 vice-chairmen, including 3 people from NTT, were elected. The new SG chairmen are listed in **Table 1**,

Table 1. Newly elected SG chairmen, and vice chairmen from Japan.

SGs	Chairmen	Vice-chairmen from Japan
TSAG (work plan)	Bruce GRACIE (Canada)	
Review committee	Yoichi MAEDA (TTC, Japan)	
SG2 (Operation)	Sherif GUINENA (Egypt)	
SG3 (Economic issues)	Seiichi TSUGAWA (KDDI, Japan)	
SG5 (Environment)	Ahmed ZEDDAM (France)	
SG9 (CATV)	Arthur WEBSTER (US)	Satoshi MIYAJI (KDDI)
SG11 (Protocols)	Wei FENG (China)	Kaoru KENYOSHI (NEC)
SG12 (Quality)	Kwame BAAH-ACHEAMFOUR (Ghana)	Akira TAKAHASHI (NTT)
SG13 (Future Networks)	Chae-Sub LEE (Korea)	Yoshinori GOTO (NTT)
SG15 (Transport Networks)	Steve TROWBRIDGE (USA)	Noriyuki ARAKI (NTT)
SG16 (Multimedia)	Yushi NAITO (Mitsubishi, Japan)	
SG17 (Security)	Arkadiy KREMER (Russia)	Koji NAKAO (KDDI)

along with the vice-chairmen elected from Japan.

From NTT, Mr. Akira Takahashi of Network Technology Laboratories was elected vice-chairman of SG12, Quality of Services, Mr. Yoshinori Goto, one of the authors and also of Network Technology Laboratories, was elected vice-chairman of SG13, Future Networks, and Mr. Noriyuki Araki of Access Service Systems Laboratories was elected vice-chairman of SG15, Access and Transport Networks.

3. Overview of other results

3.1 Approval of important Recommendations

At WTSA, ITU-T Recommendations that were controversial and difficult to approve at the SG level were discussed for approval, and six Recommendations were approved in WTSA-12. In particular, the Recommendations concerning MPLS-TP (multiprotocol label switching transport profile: technologies for highly reliable packet transport networks with the same functionalities and performance as existing transport networks through the use of MPLS), G.8113.1 and G.8113.2, were among such Recommendations. Consequently, the IANA (Internet Assigned Number Authority) assigned code points for these Recommendations. These Recommendations on MPLS-TP have been discussed intensively by ITU-T and IETF (Internet Engineering Task Force) for eight years. NTT has actively contributed to completing these Recommendations.

3.2 Establishment of Review Committee

To facilitate more efficient studies in the future study period starting in 2017, the establishment of a Review Committee (RevCom) was proposed by Japan and agreed. This group will discuss SG structure and coordinate their efforts with other standards developing organizations (SDOs) to minimize conflict of other standards with ITU-T standards. Mr. Maeda, who is the secretary general of TTC and who formerly worked for NTT, was elected as the chairman of this group.

3.3 Adoption of major resolutions

- Resolution on e-Health

The Arab region proposed a new resolution for pushing forward e-Health standardization. Although the USA opposed this for reasons involving the protection of private data of patients, the new resolution was adopted as Resolution 78 with a note recognizing the importance of privacy protection.

- Resolution on SDN

APT proposed a new resolution for SDN. Some countries such as the USA and the UK expressed views that this new resolution was not necessary because SG13 had already initiated discussions on this topic. Japan expressed the view that this new resolution would increase the visibility of SDN studies in ITU-T. As a result, this new resolution was adopted as Resolution 77.

China proposed the establishment of a new focus group for SDN. This was not accepted due to overlap with SG13. The plan for a workshop on SDN was included into this resolution in order to coordinate efforts with ITU-T and other SDOs and to advance the visibility of ITU-T SDN work.

- Resolution on e-Waste

The Arab region proposed a new resolution on e-Waste, i.e., waste electrical and electronic equipment. The UK and France opposed this proposal since Resolution 73, which addresses ICT and climate change, covers e-Waste issues, and SG5 is already studying this area. After the discussion, this proposal was agreed and adopted as Resolution 79 to express the active position of ITU-T on the standardization of e-Waste.

4. Future plans

At WTSA-12, the action plan, which includes the structure of standardization work within ITU-T in the study period from 2013 to 2016, was determined. Three people from NTT were elected as vice-chairmen of important SGs; thus, NTT will continue to have a major influence on the standardization work of ITU-T.

However, the Arab and African regions offered many comments at WTSA-12, which was held in the Middle-East, and therefore, developing countries can be expected to participate more actively in ITU-T standardization work in the near future.

From an NTT Group point of view considering the deployment of NTT solutions to developing countries, NTT would like to strive to continue participating in standardization activities in order to spread NTT technologies globally and to foster relationships with developing countries through ITU-T management team and ITU-D and ITU-R activities.

Reference

- [1] <http://www.itu.int/en/ITU-T/wtsa12/Pages/default.aspx>



Hideo Imanaka

Senior Manager, R&D Planning Department, NTT.

He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Mie University in 1985, 1987, and 2001, respectively. After joining NTT Telecommunication Network Laboratories in 1987, he engaged in research on fiber optic access network architecture and network operation process reengineering methods. From 1996 to 2003, he worked on enterprise resource planning (ERP) system integration as a consultant in the Solutions Business Division of NTT Communications. Since 2004, he has been involved in NGN standardization work at ITU-T. He was the Rapporteur of Question 1 of Study Group 13 from 2007–2010. He has also played an active role in IPTV standardization work at ITU-T. He is currently in charge of standardization strategies in the NTT Group. He received the ITU-AJ Award from the ITU Association of Japan in 2009. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Society of Instrument and Control Engineers.



Yoshinori Goto

Senior Research Engineer, Network Technology Project, NTT Network Technology Laboratories.

He received the B.E. and M.E. degrees in applied physics from Tohoku University, Miyagi, in 1992 and 1994, respectively. He joined NTT Basic Research Laboratories in 1994. He has been involved in R&D of cable television systems, IPTV, and M2M. Since 2006, he has been engaged in the standardization work for IPTV in ITU-T as a member of the IPTV Focus Group (FG-IPTV) and Global Standards Initiative (IPTV-GSI). He has also served as Rapporteur of Question 11 of ITU-T SG9, Questions 5 and 25 of ITU-T SG13, and Question 21 of ITU-T SG16. At WTSA-12, he was appointed as a vice-chairman of ITU-T SG13. He is a member of IEICE.



NTT Beijing Representative Office

Daisuke Ikegami,

NTT Beijing Representative Office

Abstract

Located in Beijing, China, the NTT Beijing Representative Office is the holding company's only Asian office outside of Japan, and it serves to support the global expansion of the NTT Group's business from an overseas viewpoint.

This article describes how the NTT Beijing Representative Office supports the global implementation of research and development results in the course of its main activities, and outlines the latest trends of NTT Group business companies in China.



1. Introduction

1.1 Overview of NTT Beijing Representative Office

The NTT Group's relationship with China dates back over 30 years, and it has evolved in various ways since 1980, when the group was still known as Nippon Telegraph and Telephone Public Corporation (Dendenkosha). In 1980, Nippon Telegraph and Telephone Public Corporation and the Chinese Ministry of Posts and Telecommunications concluded a *Memorandum of Understanding on Technology Exchange* and implemented a project to transfer Japanese crossbar switching systems to China from 1983 to 1992. Soon after, in 1985, this transfer project led to the establishment of the NTT Beijing Representative Office, which has served as the NTT Group's overseas base for the past 28 years.

The Beijing Representative Office is controlled by its shareholder, Global Business Office (**Fig. 1**) and is one of the Group's overseas offices alongside those in Washington, D.C. and San Jose, California. We explain in detail here how the Beijing Representative Office supports research and development (R&D) at NTT research laboratories through a number of activities.

1.2 Activities of Beijing Representative Office

In broad terms, the Beijing Representative Office's support of R&D can be divided into two efforts: monitoring standardization in China and facilitating the overseas implementation of NTT research laboratory technologies. For the first effort, the Beijing Representative Office is participating with observer status in the standardization activities of the China Communications Standards Association (CCSA), which is the equivalent to Japan's TTC (Telecommunication Technology Committee) and ARIB (Association of Radio Industries and Businesses) combined. Its main task is to collect information. By gathering information at the drafting stage of standardization, the Beijing Representative Office is able to quickly detect changes of direction in China's information and communications technology (ICT) field, and to provide feedback to NTT research laboratories. The Beijing Representative Office is also taking part in TC10—the CCSA's technical committee on ubiquitous network standardization—and is gathering information on the Internet of Things (IoT), which has attracted considerable attention in China in recent years.

In terms of support for the overseas implementation of NTT research laboratory technologies, the Beijing Representative Office supports the introduction of laboratory technologies, and through technological

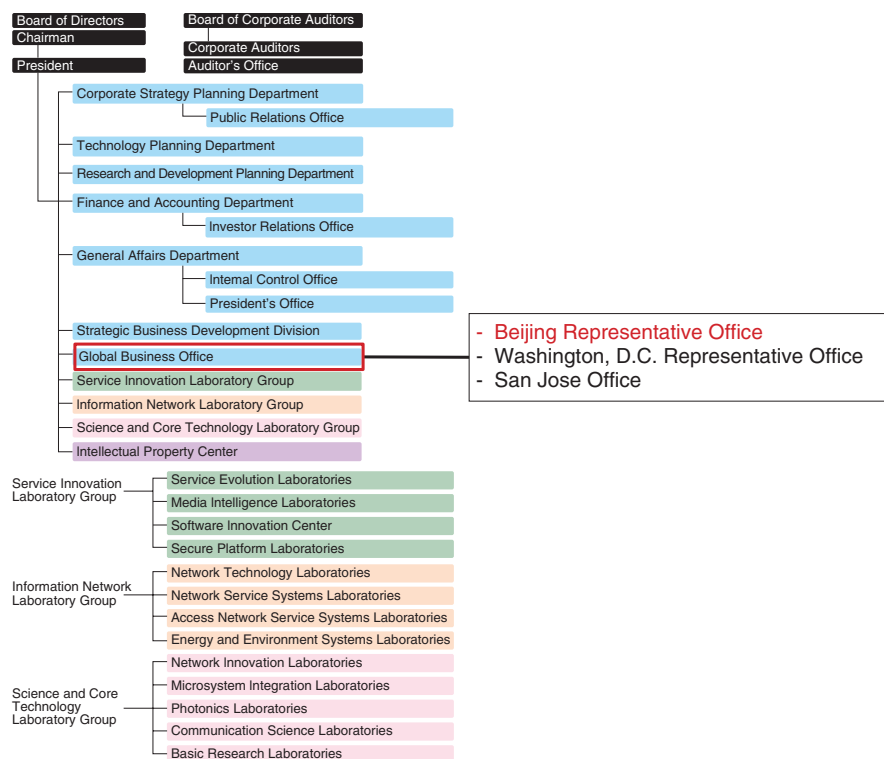


Fig. 1. The position of the Beijing Representative Office in the NTT Group.

communication between researchers, supports the development of products using laboratory technologies, mainly in Greater China.

For example, NTT has concluded a Memorandum of Understanding with the China Academy of Telecommunication Research (CATR), a Chinese government-affiliated research institution; through the exchange of opinions on areas of mutual interest, cooperation toward international standardization (not exclusive to each other's technologies) has become possible.

The representative initiatives in China by the NTT Group are shown in **Fig. 2**. In the 1990s, the NTT Group's involvement in China deepened further with its participation in projects for the Chinese postal savings system and the People's Bank of China's banking system. For example, China Union Pay is a Chinese debit card system commonly used in supermarkets and restaurants within the country, and the logo is becoming more common in Japan in electronic retail stores and major supermarkets. The company linking the China Union Pay system to the CAFIS (Credit and Finance Information System) system of Japanese banks is NTT DATA. It is probably fair to

say that the NTT Group's activities have played a role in making travel in Japan more convenient for Chinese tourists, who are subject to restrictions on the amount of foreign currency they can take out of China, and in boosting consumption.

2. Latest technological trends in China

China's ICT environment has evolved considerably in recent years. According to the Ministry of Industry and Information Technology, by the end of 2012, the number of mobile phone users had grown to about 1.11 billion, while there were about 175 million broadband users. About 80% of mobile phone users have a 2G contract, so 3G service has not yet been widely accepted. However, commercial services for TD-LTE (time-division Long Term Evolution)-based 3.9G mobile phones are expected to be launched by the end of this year. The transition to broadband is also advancing rapidly; in 2012, the number of broadband users increased by about 25 million, and the number of FTTH (fiber to the home) ports also increased by about 36 million in one year, for a total of about 260 million ports provided.

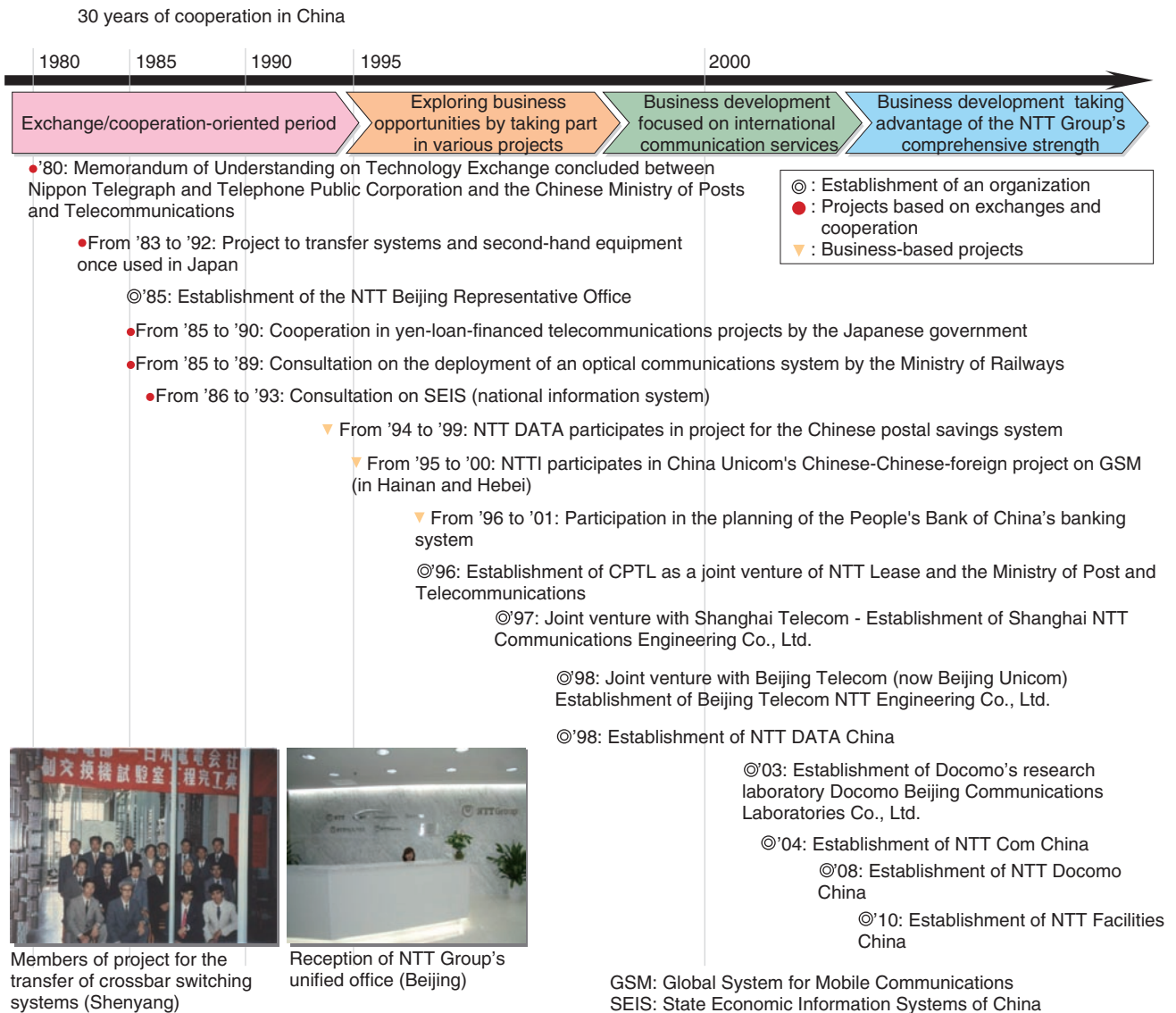


Fig. 2. The NTT Group's history in China.

The unique Chinese way of implementing instructions on directionality, which is conveyed through government policies, has played an important role in the evolution of China's ICT. In China, the central government lays out its policies every five years, determining the areas in which industrial development will be focused on. The 12th Five-Year Guideline stipulates that the seven industrial fields to be focused on from 2011 to 2015 will be energy conservation and environmental protection, next-generation information technology (IT), biotechnology, production of cutting-edge equipment, new energies, new materials, and new-energy cars. The popularization of

broadband is considered to be the focus in next-generation IT, and various measures to achieve this have been devised. In particular, national compulsory standard GB 50846-2012 (Code for design of communication engineering for FTTH in residential districts and residential buildings), enforced on April 1 of this year, stipulates in prefecture-level cities or larger (equivalent to Japanese cities) that offer FTTH, mandatory FTTH wiring for each household when new multi-unit apartments and residential buildings are constructed. The enforcement of this compulsory standard is expected to contribute to the further popularization of FTTH.

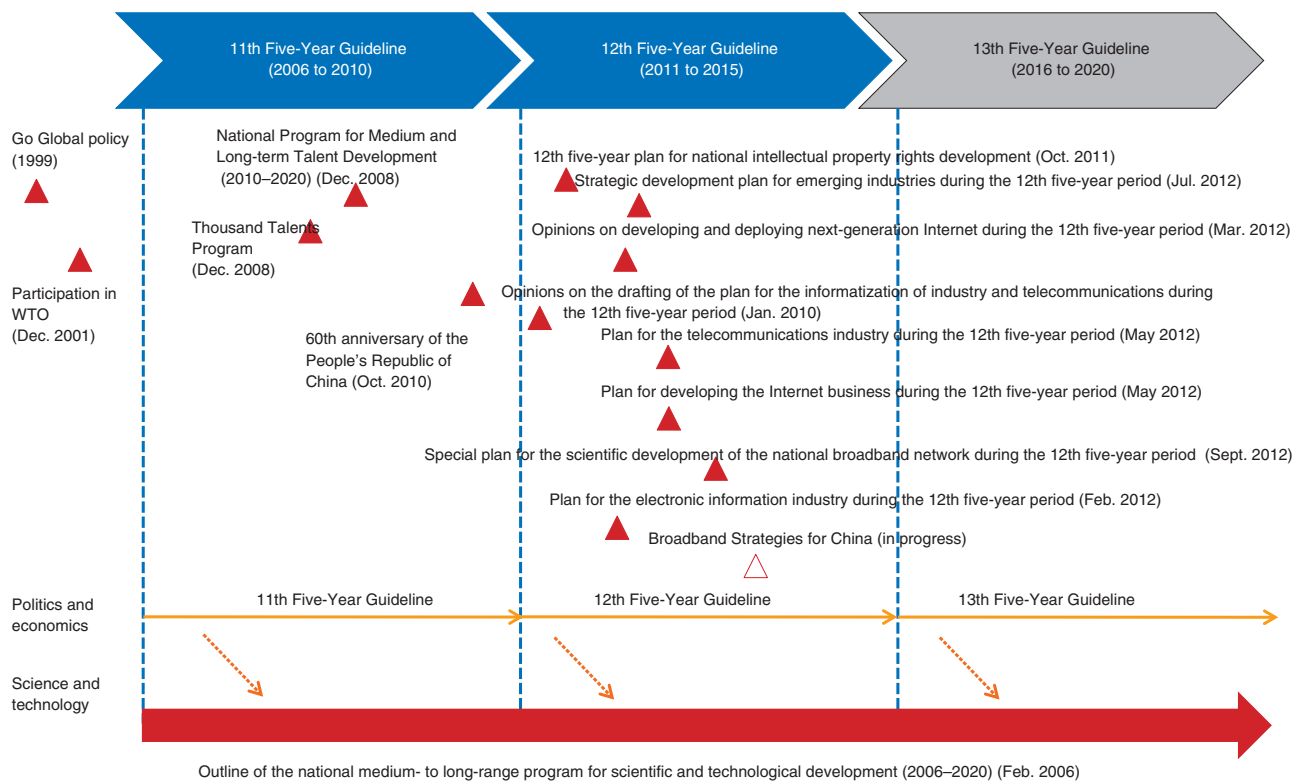


Fig. 3. Examples of Chinese policies and development plans.

As mentioned above, the central government sets key industrial fields in its Five-Year Guideline, and macro-development plans to achieve those goals are also sequentially announced. In response to these development plans, ministries, agencies, and local governments have also set numerical targets to be achieved by the end of the period of the 12th five-year plan (the end of 2015) as well as numerous guidelines and plans for conceptual attained levels. An outline of the representative policies and the various types of developmental plans is shown in **Fig. 3**.

3. Standardization activities in China

The Beijing Representative Office is participating in the standardization activities carried out by the CCSA. The following is an explanation of standards in China. In accordance with the provisions of the Standardization Law of the People's Republic of China (enforced in 1989), Chinese standards are divided into four categories: national standards, departmental standards (industry standards), local standards, and corporate standards (**Fig. 4**). Each of these standards is further divided into compulsory

standards and optional standards (recommended standards). In standardization documents, compulsory national standards are indicated with *GB* followed by a number, while optional national standards are indicated as *GB/T*. Compulsory standards for the communication industry are identified with *YD*, while optional ones are shown as *YD/T*, thus allowing one to determine the role of each established standard based on its name. The categories of Chinese technological standards are listed in **Table 1**. It should be noted that the abbreviations here all indicate Chinese terms, so have not been defined.

The Beijing Representative Office is also participating in CCSA standardization efforts related to the Internet of Things (IoT). Mentioned as a national project in the 12th Five-Year Guideline, the IoT has been attracting considerable attention in China as a field of research and has generated significant activity including the establishment of university courses focusing on it. R&D of IoT and M2M (machine-to-machine) has also been progressing in Japan, and the possibility of applying it in the fields of power grids and transportation is being explored. However, China is characterized by the considerable expectations that

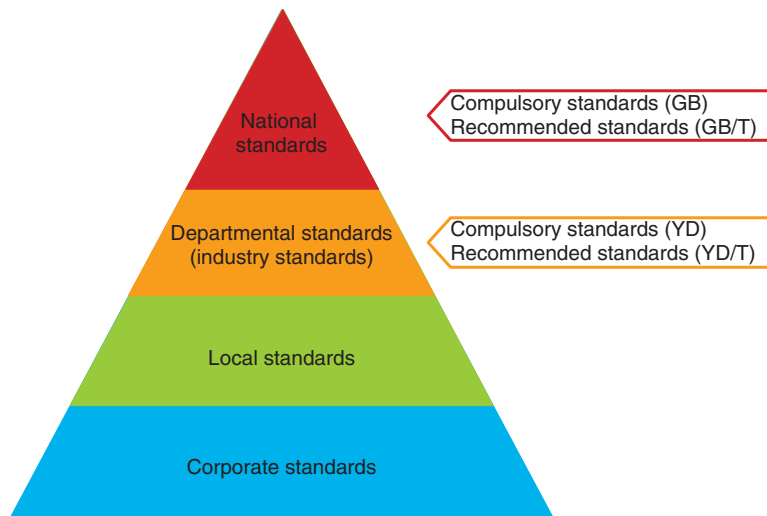


Fig. 4. Categories of standards in the Chinese standardization system.

Table 1. Categories of technological standards.

Series	Category
GB	National standards
YD	Communication industry standards
YDC	Communication standard technical documents for reference
YDB	CCSA standards*
SR	Research reports
SJ	Electronics industry standards
GY	Broadcasting standards
DL	Electricity industry standards
GJB	National military standards
OHER	Other industry standards

* Prior to 2012 it was called the Communication Standard Technical Report.

are placed on investing in fields such as agriculture and ecology in addition to those above.

4. Latest business company trends

The NTT Group’s business companies are currently expanding their operations in China, and when minor investments are included, their activities involve over 40 offices with more than 10,000 employees. Notable press releases made recently by business companies are listed in **Table 2**.

5. Future developments and challenges

As described before, the NTT Group has main-

tained a relationship with China in a variety of ways for the past 30 years. The relationship initially started for the purpose of providing technological cooperation and support, but it has gradually evolved to become more business-oriented with the establishment of offices and subsidiaries. While recent reports suggest that China’s growth is slowing down, the growth rate is said to remain high, and demand in the field of ICT is expected to keep increasing. We aim to earn the trust and appreciation of the people in China and other Asian countries, and while making efforts to achieve this, we intend to put the NTT Group’s cutting-edge ICT technologies and know-how to use in managing our business, and to contribute to the development of Asian countries and the popularization of

Table 2. Main recent press releases.

Time of the press release	Main contents	Companies concerned
March 2013	Launching of “d game” for the Chinese market	NTT DOCOMO, D2C China
January 2013	A Chinese subsidiary's general manager receives the China Economy New Leader 2012 award.	NTT Communications China
September 2012	Capital alliance with Shanghai-based Infotech Inc.	NTT DATA CHINA
May 2012	Business partnership in China's BtoC EC field	DOCOMO China/NTT DATA INSTITUTE of MANAGEMENT CONSULTING
April 2012	Capital alliance with Shanghai-based All In Finance, Inc.	NTT DATA CHINA

Japan's ICT technology.

One issue that has been brought up is the improvement of the NTT brand name power within Asia. In Japan, NTT and the dynamic loop logo are well recognized; however, in Asia they are not necessarily

recognized outside the circle of technicians. We will therefore work in the future to strengthen NTT's brand name so as to ensure that each of our group companies can conduct their business smoothly.

NTT Group unified office—short column

Christmas in China

As in the rest of the world, Christmas is an important event in China, too, but decorative illumination on the streets is not as common. However, Santa Claus costumes are very popular. Starting about two weeks before Christmas, the employees at many restaurants greet their customers with Santa hats on, and restaurants serving exclusively Chinese food are no exception. While such attire is common in China, I personally can't help feeling that it's somewhat out of place, as I fail to see the connection between Christmas—a Christian festivity—and Chinese food.



A somewhat Chinese-style Santa Claus decoration.

New Year's Day and the Chinese Spring Festival

In China, the *New Year* means the lunisolar one rather than the one on January 1. For this reason, many of the Happy New Year signs that appear on the streets after Christmas often stay there until the Spring Festival instead of being removed soon after the New Year's Day we celebrate on January 1. This year, the Spring Festival fell on February 9. The puzzling sight of Happy New Year signs well into January may be unique to China. Incidentally, a shoe store located in the basement of the Beijing Representative Office's building had a Happy New Year sign up even at the end of March. It will be interesting to see when they finally take it down!

Of China's yearly festivities, the Spring Festival is the liveliest. People migrate throughout the country in large numbers. For 40 days around the Spring Festival, when *chunyun* (a special railway schedule to deal with the needs of passengers returning home for Spring Festival holidays) is in effect, the total number of passenger journeys across the country exceeds several times China's population every year, with an estimate of 3.407 billion journeys this year. Also, it is customary to bring a variety of gifts when returning to one's hometown, which greatly boosts sales before and after the Spring Festival. China has a similar custom to Japan's tradition of giving New Year's

gifts. Here, it is called *hongbao* (red envelope). Traditionally, *hongbao* are given not only to small children, but also to parents or in-laws. Some people spend as much as a month's income on *hongbao* alone.

During the Spring Festival, the Chinese also hold year-end parties as we do in Japan. Many local companies hold lotteries with extravagant gifts such as smartphones, tablets, and microwave ovens as prizes to reward employees for their efforts in the course of the year.



A "Happy New Year" sign
—still up in March.

* Corporate social responsibility

The NTT Group's Spring Party

Upon the completion of the *NTT Group CSR* Report 2012—Our Initiatives in China*, the Chinese arm of the NTT Group held the NTT Group Beijing Spring Party on February 1, 2013. It was a successful event attended by 94 employees from 14 NTT Group companies mainly in Beijing, which included NTT Communications, Dimension Data, NTT DOCOMO, NTT DATA, and NTT Facilities. The majority of the participants were Chinese employees. The event further helped foster a sense of unity within the Group, as the facilitation by the Master of Ceremonies and the speeches by each company's representative were all conducted in Chinese.



Participants of the NTT Group Beijing Spring Party.

Daisuke Ikegami, NTT Beijing Representative Office
Hironori Nagaura, NTT DATA CHINA
Soichiro Takasugi, NTT Facilities China

External Awards

ITU-AJ Award (ICT Field Accomplishment)

Winner: Akira Takahashi, NTT Network Technology Laboratories

Date: May 17, 2013

Organization: The ITU (International Telecommunication Union) Association of Japan

For his achievements in the ICT field.

ICA Unsupervised Learning Pioneer Award 2013

Winner: Hiroshi Sawada, NTT Service Evolution Laboratories

Date: May 1, 2013

Organization: SPIE (Society of Photo-Optical Instrumentation Engineers)

In recognition of his pioneer contributions to unsupervised learning ICA (independent component analysis).