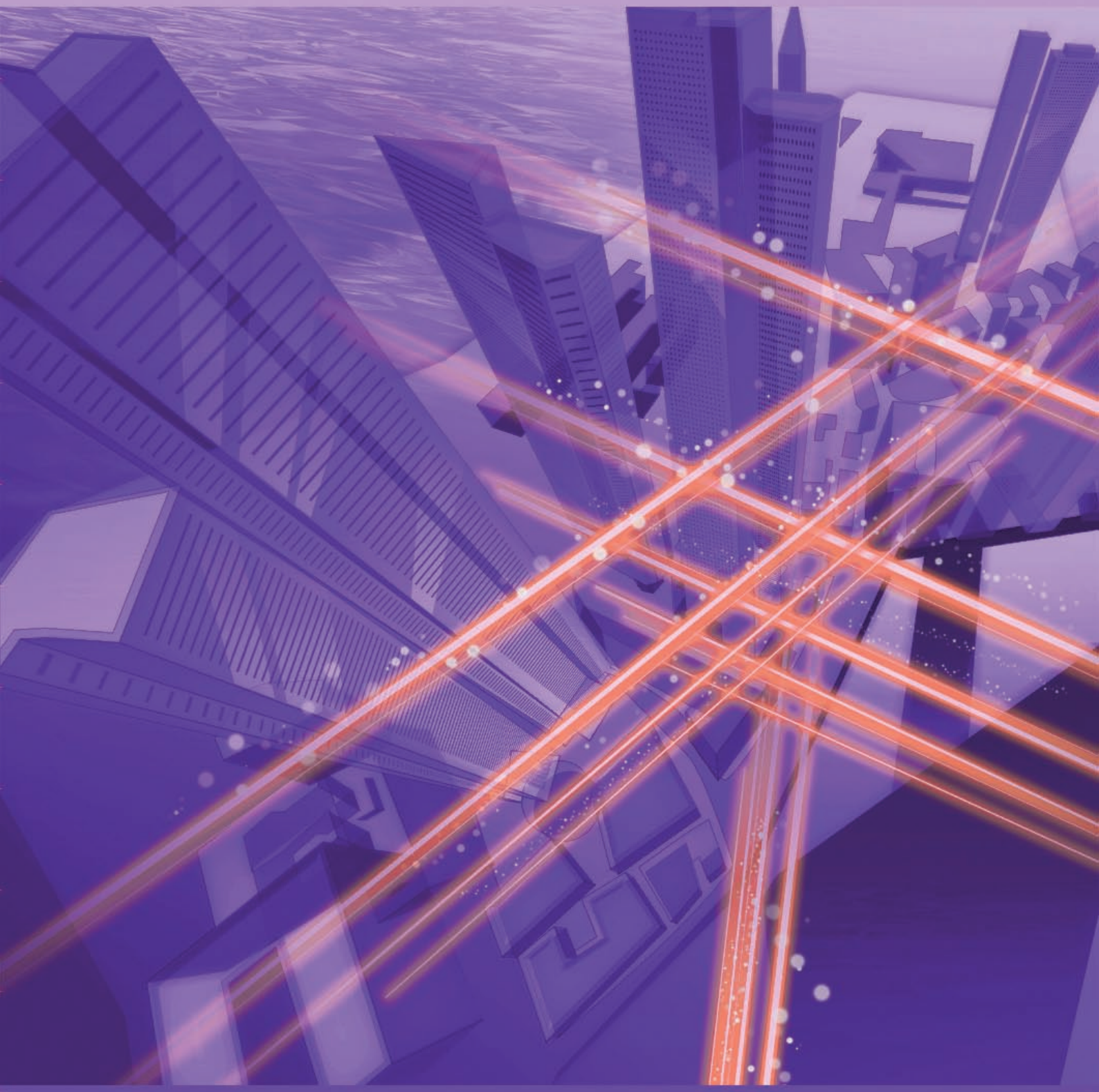
NTT Technical Review

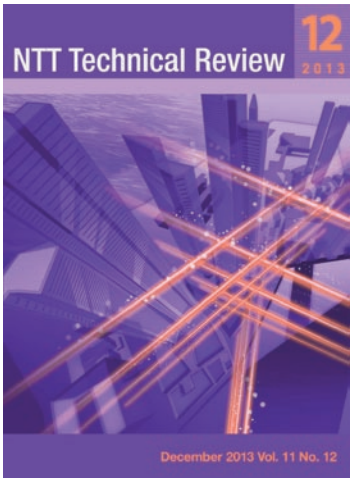
12
2013



December 2013 Vol. 11 No. 12

NTT Technical Review

December 2013 Vol. 11 No. 12



Feature Articles: Front-line of Speech, Language, and Hearing Research for Heartfelt Communications

Advanced Research in Speech, Language, and Hearing for Communication of the Future

Recent Innovations in NTT's Statistical Machine Translation

Advances in Multi-speaker Conversational Speech Recognition and Understanding

Speech Recognition Based on Unified Model of Acoustic and Language Aspects of Speech

Speaking Rhythm Extraction and Control by Non-negative Temporal Decomposition

Link between Hearing and Bodily Sensations

Regular Articles

Efficient Mining Algorithms for Large-scale Graphs

Global Standardization Activities

Trends Concerning Standardization of OpenADR

Practical Field Information about Telecommunication Technologies

Enhancing the Reliability of Aerial Iron Fittings (Span Clamps and Outdoor Wire Anchors)

Information

Report on NTT Communication Science Laboratories Open House 2013

Papers Published in Technical Journals and Conference Proceedings

Papers Published in Technical Journals and Conference Proceedings

Advanced Research in Speech, Language, and Hearing for Communication of the Future

Eisaku Maeda

Abstract

Research at NTT Communication Science Laboratories draws on both information science and human science with the aim of building a new technical infrastructure that will connect humans and information. These Feature Articles introduce new trends in the fields of speech, language, and hearing, which have a relatively long history of basic research.

Keywords: communication science, basic research, information science

1. Introduction

As we move forward in the 21st century, we are witness to the dramatic changes occurring at a truly amazing pace in the information environment that surrounds us in our daily lives. This phenomenon is clearly reflected by the transition in well-known keywords over the last ten years, for example, from *ubiquitous, grid*, and *sensor network* to *Semantic Web, Web 2.0, cloud,* and *big data*. Similarly, the information devices used to access networks have migrated from mobile phones and desktop computers to smartphones and tablets, and the range of users has expanded to include children and the elderly. These developments have also changed the face of provided services and generated a need for research and development tailored to these changes in the environment.

At NTT Communication Science Laboratories, we seek to build a new technical infrastructure that connects humans with information amid these dramatic changes in the information environment. In contrast to service development, which seeks to meet current needs, basic research aims to bring about technical innovations from a medium- and long-term viewpoint. However, as the pace of change accelerates, the strategies used to advance basic research must also change. NTT Communication Science Laboratories promotes research in a variety of scientific fields in information science and human science. These can be broadly divided into four areas: signal processing, media processing, computer intelligence, and human science (**Fig. 1**). Of particular interest here is that our successes in recent years have almost without exception combined multiple fields and technologies in a synchronized and skillful manner. This outcome can be seen in both scientific fields and service development. It is safe to say that each and every researcher in the upcoming era will need to have a good background in multiple fields.

2. Cultivating trees that bear fruit

Minor paradigm shifts or the encountering of problems often become the seeds for new research, and if those seeds are given water, they will eventually sprout. If the resulting buds are then given fertilizer and exposed to sunlight, they will grow into trees, and on those trees, flowers representing patents, papers, and other achievements will bloom. Although flowers cannot normally be eaten, they give forth fruit that can be picked. However, this harvested fruit cannot always be eaten in its original state. There is hard, unripe fruit, sour fruit, and even poisonous fruit. There is also some fruit that must be prepared and cooked in various ways before being eaten while some fruit can be stored away and preserved for later

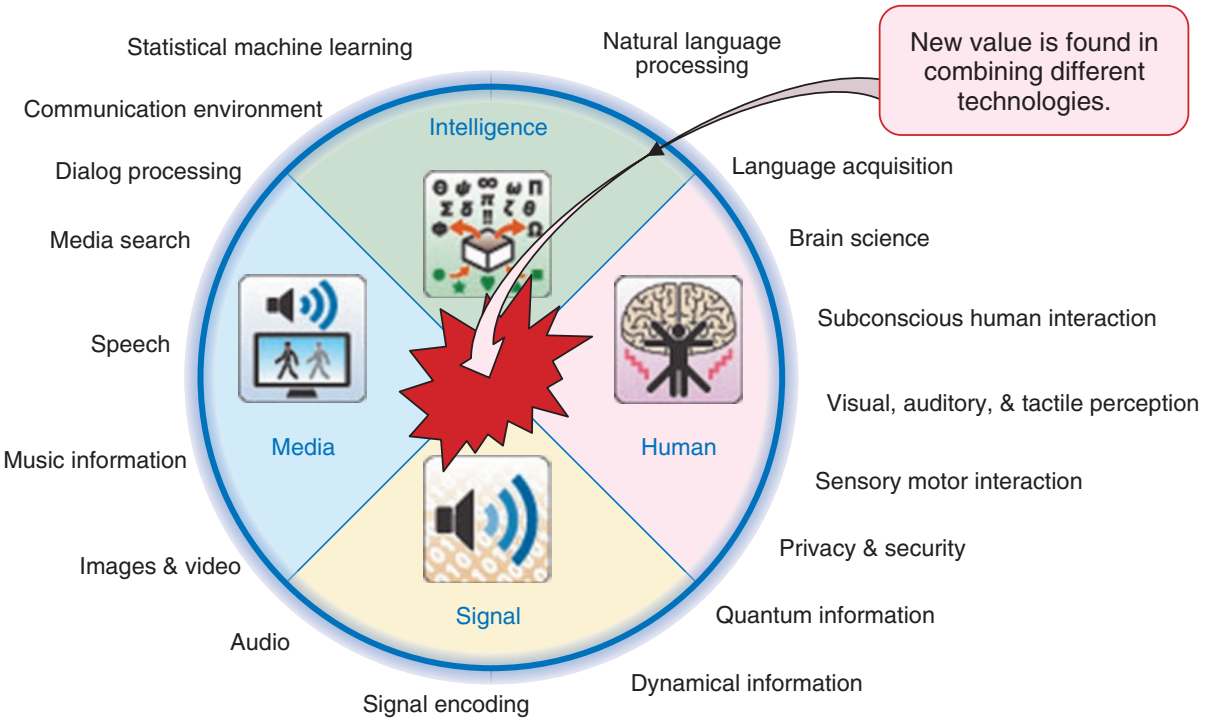


Fig. 1. Research fields at NTT Communication Science Laboratories.

use.

In any case, just as fruit will eventually nourish people in one shape or another, we can treat the results of research as technology that will eventually serve a useful purpose in society. Here, trees signify research, and *cultivating trees that bear fruit* is the most important role of basic research. It's been more than 20 years since the founding of NTT Communication Science Laboratories, and our technologies that have found a place in society have been increasing slowly but surely. If we look at examples of our successes in recent years in areas like media search, speech recognition, reverberation control, question answering, statistical translation, and *texture* information science, we can see that a period of about ten years is needed after sowing the seeds before any fruit will be ready to consume.

3. Advanced research in speech, language, and hearing

The five articles concerning speech, language, and hearing in these Feature Articles report on recent research achievements of NTT Communication Science Laboratories. These achievements hold various positions in the research scenarios based on the fruit analogy described above.

Machine translation technology entitled "Recent Innovations in NTT's Statistical Machine Translation" [1] represents a genuine era of practical application after a long research history spanning more than 30 years. The accumulation of language resources and know-how through years of research as well as recent technical innovations lie behind this innovative development period.

The two articles entitled "Advances in Multi-speaker Conversational Speech Recognition and Understanding" [2] and "Speech Recognition Based on Unified Model of Acoustic and Language Aspects of Speech" [3] introduce the latest trends in speech recognition technology. We are now entering an era of practical speech recognition technology that will enable speech recognition to be used, for example, in preparing the minutes of proceedings in Japan's National Diet. The speech recognition field, however, still has some issues that must be solved depending on the usage environment and application. Here as well, the combination of multiple technologies will give rise to new technologies that have a competitive advantage, and the germination of these technologies

has already started.

"Speaking Rhythm Extraction and Control by Non-negative Temporal Decomposition" [4] is an achievement born of research into the mechanism of human speaking. The human voice is generated as a sound originating in the vibration of speech organs, and we are working on fascinating developments by combining that process with information-science technologies related to speech processing.

"Link between Hearing and Bodily Sensations" [5] introduces research that was the first in the world to unravel the relationship between human bodily sensations and the sense of hearing. The so-called flowers are finally blooming because of this research, and we look forward to seeing what kinds of fruit these flowers will bring forth.

References

[1] M. Nagata, K. Sudoh, J. Suzuki, Y. Akiba, T. Hirao, and H. Tsukada, "Recent Innovations in NTT's Statistical Machine Translation," NTT Technical Review, Vol. 11, No. 12, 2013. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201312fa2.html

[2] T. Hori, S. Araki, T. Nakatani, and A. Nakamura, "Advances in Multi-speaker Conversational Speech Recognition and Understanding," NTT Technical Review, Vol. 11, No. 12, 2013. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201312fa3.html

[3] Y. Kubo, A. Ogawa, T. Hori, and A. Nakamura, "Speech Recognition Based on Unified Model of Acoustic and Language Aspects of Speech," NTT Technical Review, Vol. 11, No. 12, 2013. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201312fa4.html

[4] S. Hiroya, "Speaking Rhythm Extraction and Control by Non-negative Temporal Decomposition," NTT Technical Review, Vol. 11, No. 12, 2013. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201312fa5.html

[5] N. Kitagawa, "Link between Hearing and Bodily Sensations," NTT Technical Review, Vol. 11, No. 12, 2013. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201312fa6.html

Eisaku Maeda

Director, NTT Communication Science Laboratories.

He received the B.E. and M.E. degrees in biological science and the Ph.D. degree in mathematical engineering from the University of Tokyo in 1984, 1986, and 1993, respectively. He joined NTT in 1986. He was a guest researcher at the University of Cambridge, UK, during 1996–1997. His research interests are in statistical machine learning, intelligence integration, and bioinformatics. He is a senior member of IEEE and a member of the Institute of Electronics, Information and Communication Engineers and the Information Processing Society of Japan.

Recent Innovations in NTT's Statistical Machine Translation

Masaaki Nagata, Katsuhito Sudoh, Jun Suzuki, Yasuhiro Akiba, Tsutomu Hirao, and Hajime Tsukada

Abstract

English and Japanese have very different word orders, and they are probably one of the most difficult language pairs to translate. We developed a new method of translating English to Japanese that takes advantage of the head-final linguistic nature of Japanese. It first changes the word order in an English sentence into that of a Japanese sentence and then translates the reordered English sentence into Japanese. We found that our method dramatically improved the accuracy of English-to-Japanese translation. We also found that the method is highly effective for Chinese-to-Japanese translation.

Keywords: statistical machine translation, language distance, preordering

1. Introduction

Machine translation is a technology to translate one language into another language by computer. Research on machine translation started in the 1950s, making it almost as old as the computer itself, and many different types of machine translation systems have been developed over the years.

Many web pages nowadays are written in various foreign languages such as Chinese, Korean, and Arabic due to the advancement of Internet technologies. Multinational companies must translate their manuals and product information quickly and accurately into the local languages. It may be said that it is a fundamental desire for human beings to break through language barriers in order to communicate and exchange knowledge. However, previous machine translation systems have not been capable of satisfying the various translations needs of users.

2. From rule-based translation to statistical translation

Previous machine translation systems required the development of large-scale translation rules and bilingual dictionaries for each language pair. This is a labor-intensive task that requires the efforts of dozens of specialists over several years. This kind of machine translation approach is called *rule-based translation*.

It is often said that rule-based translation has reached its limit in accuracy and is difficult to improve further. *Statistical machine translation* is proposed as an alternative to rule-based translation. It automatically learns statistical models, which are equivalent to translation rules and bilingual dictionaries, from a large number of bilingual sentences—on the order of several hundred thousand to several million. It is an emerging technology in which the ultimate goal is to develop a machine translation system for new language pairs or new domains at low cost and in a short period of time. An outline of the statistical machine translation process is shown in **Fig. 1**.

In around 1990, researchers at IBM proposed a machine translation system between French and English using the Canadian Hansard, the transcripts of parliamentary debates. This was the first attempt to use statistical machine translation. In the 2000s, the accuracy of statistical machine translation reached a level of practical use for language pairs with similar word orders, for example French and English, by applying *phrase-based translation*, in which the translation unit changed from words to phrases.

In around 2005, the translation accuracy for language pairs with larger word differences such as

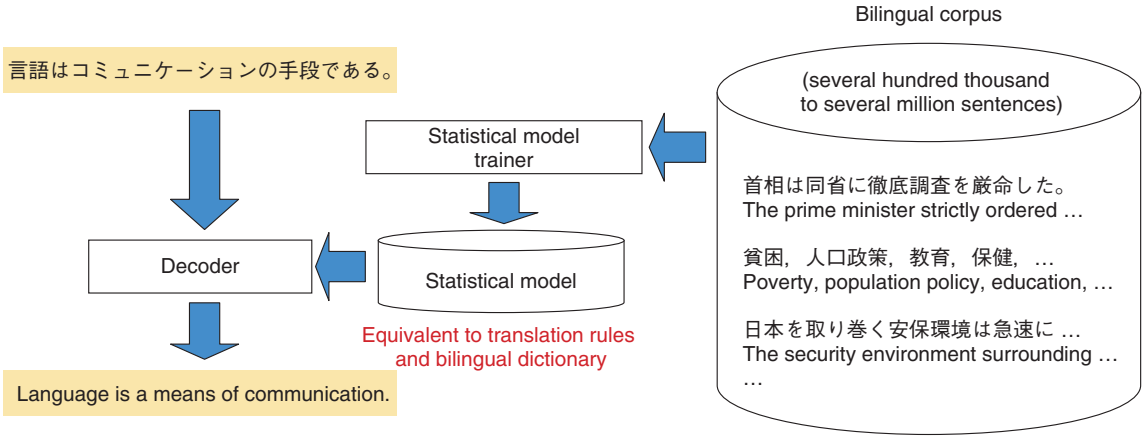


Fig. 1. Outline of statistical machine translation.

Chinese and English were improved by using *tree-based translation*, which uses syntactic theory or the hierarchical structure of either the source or target sentence. The accuracy of statistical machine translation turned out to be higher than that of rule-based translation, not only for language pairs with similar word orders such as French and English, but also for those with relatively different word orders such as Chinese and English. However, statistical machine translation could not outperform rule-based translation for language pairs with highly different word orders such as Japanese and English.

3. Preordering for translation

The idea of *preordering*, in which the word order of the source language sentence is rearranged into that of the target language sentence before translation, was first presented in the early 2000s. It started to gain attention from major research institutes such as Google, Microsoft, IBM, and NTT as a promising technology to overcome word order difference in around 2010. Preordering involves the use of reordering rules in order to obtain the word order of the target sentence. These rules are applied to the syntactic structure obtained by parsing the source language sentence. Reordering rules are usually created manually, although some methods exist for learning them automatically from a bilingual corpus with automatic word alignments.

With respect to reordering, NTT has focused on the head-final nature of the Japanese syntactic structure and has proposed a preordering method for English-to-Japanese translation called head finalization that uses only one rule: *move the syntactic head to the end of the constituent* [1]. We found that it dramatically improves the accuracy of English-to-Japanese translation. In 2011, the joint team of NTT and the University of Tokyo ranked first in the evaluation of an NTCIR-9 (NII (National Institute of Informatics) Testbeds and Community for Information access Research, 9th meeting) patent translation task by combining the University of Tokyo's accurate English parser, Enju, with NTT's preordering technique using head finalization. This was the first time ever that, by human evaluation, the accuracy of statistical machine translation outperformed that of rule-based translation in an English-to-Japanese translation task [2], [3].

4. Preordering method based on head-final order in Japanese

An outline of the preordering method based on the head-final property of Japanese is shown in **Fig. 2**. A phrase is a constituent of a sentence, and its *head* is a word that determines the grammatical role of the phrase in a sentence. For example, a preposition is the head of a prepositional phrase. In other words, in the dependency relation, which all Japanese students learn in Japanese class in elementary school, the word that is modified is the head. In the Japanese pattern of dependency, the dependency always goes from left to right; that is, modified words are always at the sentence-end side of the words that modify them. This is the head-final property of Japanese.

In fact, Japanese is a strictly head-final language, which is very rare in the world. In general, as shown

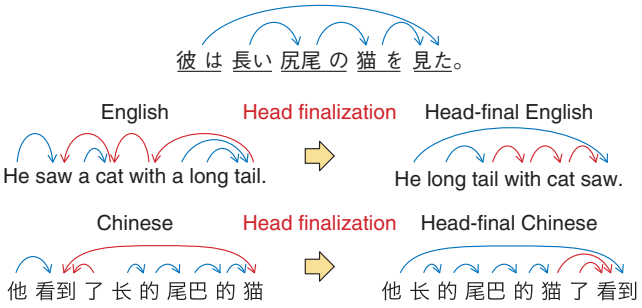


Fig. 2. Preordering method based on head-final property of Japanese.

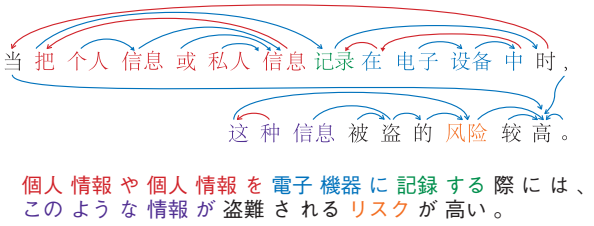


Fig. 3. Example of a dependency structure for a Chinese patent sentence.

in the English and Chinese examples in Fig. 2, dependency goes both from left to right and from right to left. The verb (saw) in the English example is modified by the subject from the left (He) and by the object from the right (cat). With respect to the two nouns, the adjective modifies one of the nouns (tail) from the left, and the prepositional phrase modifies the other noun (cat) from the right. In the Chinese example, with respect to the verb, the subject is on the left and the object on the right, but the modifications of both nouns go from left to right.

Because of this head-final property of Japanese, if we reorder the words of the source language sentence so that its dependency always goes from left to right, the resulting word order is the same as its Japanese translation. This is the basic idea of preordering based on head finalization. If the word order of a source language sentence is the same as that of the target sentence, the remaining task is word-to-word translation, which can be solved accurately by statistical machine translation. Since the head finalization only uses the linguistic properties of the target language, it can be applied to translations from any language into Japanese as long as we have a method to obtain the syntactic structure of the source language.

Translating Japanese into other languages is more difficult than translating other languages into Japanese because for each dependency relation in the syntactic structure of the source Japanese sentence, we have to decide whether to keep its direction or not based on its grammatical role in the target language.

5. Multilingual translation of technical documents

We built a statistical machine translation system to translate patent documents from English, Chinese, and Korean into Japanese in order to verify the feasibility of statistical machine translation from foreign languages into Japanese based on the proposed preordering technique.

In the patent domain, there are *patent families*; this refers to a set of patents for one invention that is applied to different countries that share priority claims to the same patent. Patent documents in a patent family are not a perfect translation of each other, but they include many sentences that are translations of sentences in other documents in the same family. Therefore, we can extract a large-scale bilingual corpus by mining patent families. We prepared three bilingual corpora, English-Japanese (about 4 million sentences), Chinese-Japanese (about 8 million), and Korean-Japanese (about 2 million), from patent documents filed in Japan, the U.S., China, and Korea from 2004 to 2012. As far as we know, the Chinese-Japanese and Korean-Japanese patent corpora are each one of the largest in their language pairs.

To apply the proposed preordering method based on head finalization, we need a technology for parsing the syntactic structure of the source language sentence accurately. We made a set of training data with manually annotated syntactic structures for English (40,000 sentences from news articles and 10,000 sentences from patent documents), and for Chinese (50,000 sentences from news articles and 20,000 sentences from patent documents). We then made a dependency parser for English and Chinese using a semi-supervised learning technique developed by NTT in 2009, which achieved the best published accuracies in international benchmark data for English and Czech dependency parsing [4].

An example of a dependency structure of a Chinese patent sentence is shown in **Fig. 3**, and an example of its Chinese-to-Japanese translation is shown in **Fig. 4**.

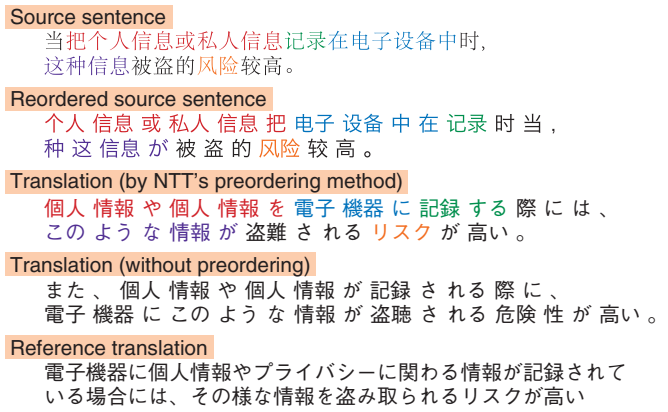


Fig. 4. Example of Chinese-to-Japanese translation of patent sentence.

In general, sentences in patents are long and have complicated dependency structures. However, we found that with head finalization, we could generate a target Japanese sentence that accurately reflected the dependencies in the source Chinese sentence if we could parse the dependency of the Chinese sentence correctly. It should be noted that we did not have to apply preordering in the Korean-to-Japanese translation because the word order in Korean is almost the same as that in Japanese.

6. Automatic evaluation of translation accuracy

Finally, we briefly explain the automatic evaluation of translation accuracy. Objective evaluation of machine translation accuracy is very difficult. There are many correct translations for a sentence, and it is a subjective decision whether to focus on word translation errors or word order errors. An automatic evaluation measure for translation called BLEU (BiLingual Evaluation Understudy), brought a revolutionary change and accelerated the research on machine translation when it was presented in the 1990s. In a sense, the effect was similar to the invention of the instrument for measuring the taste of rice, which activated the competition between rice producing areas and prompted efforts to improve the various breeds. One problem with BLEU, however, is that it does not agree with human evaluations of translations between English and Japanese. The correlation between human evaluation and automatic evaluation by BLEU for the Japanese-to-English translation of a patent translation task in NTCIR-7, which was held in 2008, is shown in **Fig. 5**.

To solve this problem, we proposed a novel automatic evaluation measure called RIBES (Rank-based Intuitive Bilingual Evaluation Score) and released it to the public as open source software [5], [6]. It focuses more on the degree of word order agreement between translation results and reference translations. RIBES was adopted as one of the official evaluation measures at the previously mentioned NTCIR-9, and organizers of the workshop found that it had higher agreement with human evaluation than BLEU in English-to-Japanese, Japanese-to-English, and Chinese-to-English translation tasks [2].

7. Practical application of machine translation

In the translation of technical documents such as patents, manuals, and scientific journals, it is very important to transfer the objective and logical meaning of the content, that is, to accurately map the modifier-modified relations from the source language to the target language. We think that statistical machine translation from foreign languages into Japanese has reached the level of practical use for domains such as patents, in which we can collect a bilingual corpus of more than one million sentences.

For translation from Japanese to English, statistical machine translation is yet to outperform rule-based translation, although the difference in accuracy is getting increasingly smaller. Future tasks include extending the application domains from technical documents to business documents and to spoken languages, as well as improving the translation from Japanese to foreign languages.

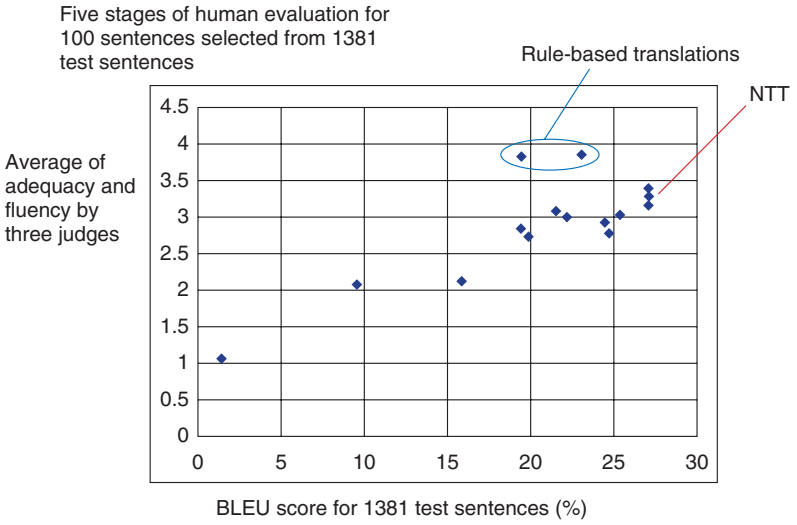


Fig. 5. Correlation between human evaluation and BLEU in NTCIR-7 Japanese-to-English patent translation task (2008).

References

[1] H. Isozaki, K. Sudoh, H. Tsukada, and K. Duh, "HPSG-Based Preprocessing for English-to-Japanese Translation," Journal of ACM Trans. on Asian Language Information Processing (TALIP), Vol. 11, No. 3, 2012.

[2] I. Goto, B. Lu, K. P. Chow, E. Sumita, and B. K. Tsou, "Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop," Proc. of NTCIR-9 Workshop Meeting, pp. 559–578, Tokyo, Japan, 2011.

[3] K. Sudoh, K. Duh, H. Tsukada, M. Nagata, X. Wu, T. Matsuzaki, and J. Tsujii, "NTT-UT Statistical Machine Translation in NTCIR-9 PatentMT," Proc. of NTCIR-9 Workshop Meeting, pp. 585–592, Tokyo, Japan, 2011.

[4] J. Suzuki, H. Isozaki, X. Carreras, and M. Collins, "An Empirical Study of Semi-supervised Structured Conditional Models for Dependency Parsing," Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 551–560, Suntec, Singapore.

[5] RIBES: Rank-based Intuitive Bilingual Evaluation Score. http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html

[6] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada, "Automatic Evaluation of Translation Quality for Distant Language Pairs," Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 944–952, Cambridge, MA, USA.

Masaaki Nagata

Senior Distinguished Researcher, Group Leader, 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. He joined NTT in 1987. He was with Advance Telecommunications Research Institutes International (ATR) Interpreting Telephony Research Laboratories, Kyoto, from 1989 to 1993. He was a visiting researcher at AT&T Laboratories Research from 1999 to 2000. 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 (IEICE), the Information Processing Society of Japan (IPSJ), the Japanese Society for Artificial Intelligence (JSAI), the Association for Natural Language Processing (ANLP), and the Association for Computational Linguistics (ACL).

Katsuhito Sudoh

Research Scientist, NTT Communication Science Laboratories.

He received the B.Eng. and M.Inf. degrees from Kyoto University in 2000 and 2002. He joined NTT in 2002 and studied spoken dialogue systems and spoken language processing. He is currently working on statistical machine translation. He is a member of IPSJ, ANLP, ACL, and the Acoustic Society of Japan (ASJ).

Jun Suzuki

Senior Research Scientist, NTT Communication Science Laboratories.

He received the B.Sc. degree in mathematics and M.Eng. degree in computer science from Keio University, Kanagawa, in 1999 and 2001 and the Ph.D. degree in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology in 2005. He joined NTT Communication Science Laboratories in 2001. He is currently studying machine learning and natural language processing areas including kernel methods, supervised and semi-supervised learning, question answering, machine translation, and natural language parsing. During 2008–2009, he stayed at MIT CSAIL in Boston as a visiting researcher to develop the high-performance dependency parser. Since September 2013, he has been a member of the editorial board of ANLP Journal. He is a member of IPSJ, ANLP, and ACL.

Yasuhiro Akiba

Senior Research Scientist at NTT Communication Science Laboratories and Senior Research Engineer at NTT Media Intelligence Laboratories.

He received the B.Sc. and M.Sc. degrees in mathematics from Waseda University, Tokyo, in 1988 and 1990 and the Ph.D. degree in informatics from Kyoto University in 2005. He joined NTT in 1990. He was with the ATR Spoken Language Translation Research Laboratories as a Senior Researcher from October 2000 to March 2005. His research interests include machine learning, knowledge acquisition, natural language learning, machine translation, and automatic evaluation. He received the Best Paper Award at the 9th Annual Conference of the Japan Society for Artificial Intelligence in 1995 and the CV Ramamoorthy Best Paper Award of the 12th IEEE International Conference on Tools with Artificial Intelligence in 2000. From April 1999 to March 2001, he was a member of the editorial board of IPSJ Magazine. He is a member of IPSJ and ANLP.

Tsutomu Hirao

Research Scientist, NTT Communication Science Laboratories.

He received the B.E. degree from Kansai University, Osaka, in 1995, and the M.E. and Ph.D. degrees in engineering from Nara Institute of Science and Technology in 1997 and 2002. He joined NTT Communication Science Laboratories in 2000. His current research interests include Natural Language Processing and Machine Learning. He is a member of IPSJ, ANLP, and ACL.

Hajime Tsukada

Senior Research Scientist, NTT Communication Science Laboratories.

He received the B.S. and M.S. degrees in information science from Tokyo Institute of Technology in 1987 and 1989. He joined NTT Human Interface Laboratories in 1989. In 1997, he joined ATR Interpreting Telecommunications Research Laboratories, and from 1998 to 1999, he was a visiting researcher at AT&T Laboratories Research. Since 2003, he has been with NTT Communication Science Laboratories. His research interests include statistical machine translation as well as speech and language processing. He is a member of IEICE, JSAI, ANLP, ACL, and ASJ.

Advances in Multi-speaker Conversational Speech Recognition and Understanding

Takaaki Hori, Shoko Araki, Tomohiro Nakatani, and Atsushi Nakamura

Abstract

Opportunities have been increasing in recent years for ordinary people to use speech recognition technology. For example, we can easily operate smartphones using voice commands. However, attempts to construct a device that can recognize human conversation have produced unsatisfactory results in terms of accuracy and usability because current technology is not designed for this purpose. At NTT Communication Science Laboratories, our goal is to create a new technology for multi-speaker conversational speech recognition and understanding. In this article, we review the technology we have developed and present our meeting analysis system that can accurately recognize *who spoke when, what, to whom, and how* in meeting situations.

Keywords: multi-speaker, speech recognition, diarization

1. Introduction

A meeting is a basic human activity in which a group of people share information, present opinions, and make decisions. In formal meetings, it is standard for one person to take minutes. However, it often happens that certain important details are forgotten and therefore not recorded in the minutes. Moreover, meetings are not always easy to control, and this sometimes makes it difficult to achieve the objectives of the meeting. The participants may also be ill-informed, which can lead to misunderstandings or disagreements. Consequently, technology that is capable of automatically recognizing and understanding speech used in meetings has been attracting increasing attention [1], [2] as a way to overcome such problems.

Today, speech recognition technology is widely used in many applications such as the operation of smartphones using voice commands. If we speak clearly into such a device, the spoken words can be recognized correctly and the command executed as intended. However, when we try to construct a device that can be applied to recognize conversations in meetings, as many as half of the words are not recognized correctly. This is because speech signals are often degraded by background noise and the voices of other participants, and conversational speech itself involves a wide variety of acoustic and linguistic patterns compared with speech directed at a device. As a result, the speech recognition accuracy deteriorates significantly. At NTT Communication Science Laboratories, we are working hard to develop meeting speech recognition technology that can solve these problems.

However, another problem is that even if a speech recognizer achieves 100 percent accuracy for a meeting, no information about the meeting will be provided except for the spoken word sequence in a text format. This means that we can understand what words were spoken in the meeting but not who spoke when, to whom, and in what manner, which are all important pieces of information if we are to understand any meeting. In our research group, we are also

Fig. 1. Image of meeting captured by camera and microphone array.

studying meeting analysis technology that will enable us to understand a meeting in its entirety [2]. Our aim is to create a system that simultaneously obtains verbal information by speech recognition and nonverbal information by audio-visual scene analysis.

We have already developed a prototype system for meeting analysis, which we designed to evaluate and demonstrate our proposed techniques. The first version of the system visualized a meeting based on nonverbal information, where the system recognized *who spoke when and to whom* and estimated the visual focus of attention using a microphone array and an omnidirectional camera [1]. We then extended the system to recognize both verbal and nonverbal information by incorporating our meeting speech recognition technology [2]. We have already shown that the system can both create draft meeting minutes and assist meeting participants with functions for looking back at past utterances and accessing information related to the words spoken during the meeting.

In this article, we review the meeting speech recognition and understanding technology we have developed. In section 2, we describe our attempts to improve meeting speech recognition. In section 3, we present our meeting analysis system that accurately recognizes *who spoke when, what, to whom, and how*. We conclude the article and touch on future work in section 4.

2. Recognition of meeting speech

2.1 Problems with meeting-speech recognition

We consider an ordinary meeting room as shown in **Fig. 1**, where four meeting participants freely discuss various topics, and all the utterances are recorded with a microphone placed at the center of the table. However, speech recognition is not easy in this situation, and the recognition result will include many errors. There are two reasons for this problem, as described below.

(1) Conversation oblivious to microphones

In face-to-face meetings, having participants wear a microphone or placing a microphone directly in front of each participant is not the preferred approach because it severely restricts the movement of the participants. Therefore, microphones should be located further away from each participant. However, this results in interference from acoustic noise and reverberation. Moreover, in settings where participants engage in informal and relaxed conversation, the utterances of two or more speakers often overlap. These factors significantly degrade speech recognition accuracy.

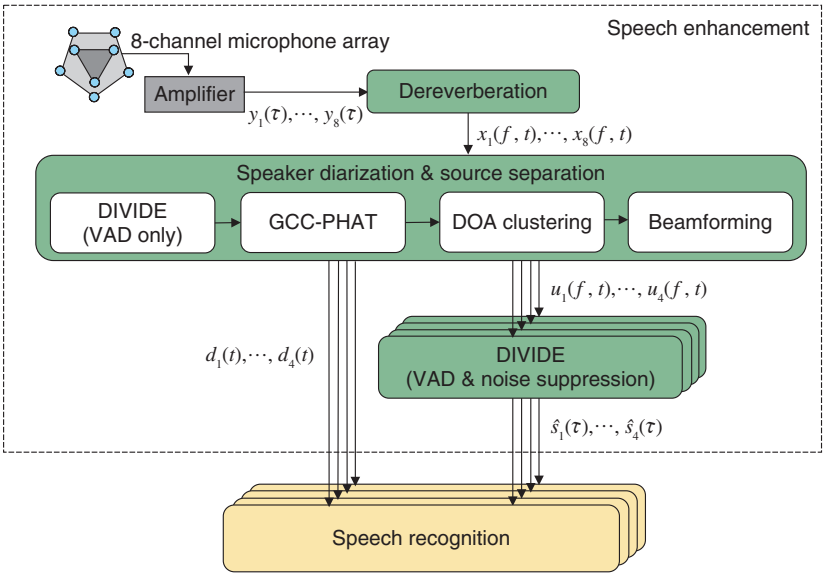


Fig. 2. Speech enhancement system.

(2) Acoustic and linguistic variety of spontaneous speech

In private or informal meetings, people rarely speak formally and clearly. From acoustic and linguistic points of view, the utterances are fully spontaneous and therefore tend to include ambiguous pronunciations, abbreviations, and dialectal and emotional expressions. Consequently, a wide variety of speech patterns exist even for words that have the same dictionary pronunciation. These patterns change greatly depending on the speaker, the speaking style, and the topic. This aspect of spontaneous speech is also a crucial problem that degrades recognition accuracy.

2.2 Solutions to problems

To solve these problems, we first worked on speech enhancement to improve the quality of speech signals in a meeting and proposed effective techniques based on a microphone array [2]. An overview of the speech enhancement method used in our meeting recognition system is shown in **Fig. 2**. The speech enhancement process consists of three phases: dereverberation, speaker diarization/source separation, and noise suppression.

(1) The dereverberation phase transforms the eight-channel microphone signals $y_1(\tau)\cdots y_8(\tau)$ into the time-frequency domain, removes reverberation components from the complex spectral sequences $y_1(f, t)\cdots y_8(f, t)$ of the microphone signals using multi-channel linear prediction, and outputs eight-channel dereverberated spectral sequences $x_1(f, t)\cdots x_8(f, t)$.

(2) The speaker diarization/source separation phase detects active speakers based on direction of arrival (DOA) information, which is estimated by applying the generalized cross-correlation method with phase transform (GCC-PHAT) [3] to the dereverberated speech spectral sequences $x_1(f, t)\cdots x_8(f, t)$. In our system, the diarization result is obtained by clustering the DOAs at each frame, and the utterance period for each speaker n is output as the binary speaker diarization results $d_1(t)\cdots d_4(t)$, where $d_n(t) = 1$ (or 0) indicates that speaker n is speaking (or silent) at frame t. Subsequently, source separation is performed to separate overlapping speech into speaker-dependent channels, where a null-beamforming approach is employed because it does not produce nonlinear artifacts that have detrimental effects on speech recognition. The beamformer coefficients for each speaker are estimated by leveraging the diarization result $d_n(t)$ [4].

(3) The noise suppression phase suppresses the noise components contained in each separated spectrum $u_n(f, t)$, where we use Dynamic Integration of Voice Identification and DE-noising (DIVIDE) [5], [6]. DIVIDE reduces only the noise component of the original signals by using the online estimation of the speech and noise components, and then it outputs time-domain enhanced speech signals $\hat{s}_n(\tau)$.

After the speech enhancement, speech recognition is performed by using the enhanced speech signals $\hat{s}_n(\tau)$, in which the diarization results $d_n(t)$ are also used to validate whether or not each recognized word is actually spoken by the participant associated with the separated channel.

We recently improved the speech enhancement further by using DOLPHIN [7], which extracts the target speech more clearly based on the acoustic patterns of speech in the time and frequency domains. Although this method does not currently work with online processing, we confirmed that it yielded a large gain in recognition accuracy of meeting speech.

2.3 Automatic speech recognition module

Next we present a brief overview of the automatic speech recognition (ASR) module that we designed for transcribing meeting speech. The module is based on SOLON [8], a speech recognizer that employs weighted finite-state transducers (WFSTs). SOLON employs an acoustic model consisting of a set of hidden Markov models (HMMs), a pronunciation lexicon, and language models represented as WFSTs that can be combined *on the fly* (i.e., as quickly as necessary) during decoding. The decoder efficiently finds the best scoring hypothesis in a search space organized with the given WFSTs.

The input signal to the ASR module is spontaneous speech uttered by meeting participants, recorded with distant microphones, and enhanced by the audio processing techniques shown in Fig. 2. In general, it is effective to use a large amount of meeting speech data and their transcriptions to train acoustic and language models. However, there are no available Japanese data recorded under similar conditions, and it is very costly to collect new meeting data. Therefore, we prepared only a small amount of matched-condition data and used them to adapt the acoustic and language models.

The acoustic model is a set of state-shared triphone HMMs, where each triphone (a sequence of three phonemes) is modeled as a left-to-right HMM with three states, and each shared state has a Gaussian mixture output distribution. First, initial HMMs are trained with a large corpus of clean speech data recorded via a close-talking microphone. The parameters of the initial HMMs are then estimated by discriminative training based on a differenced maximum mutual information (dMMI) criterion [9] to reduce recognition errors.

Next, the initial HMMs are adapted with a small amount of real meeting data, which were recorded and enhanced with our meeting recognition system in advance. That is, the data were recorded with an 8-channel microphone array, and then dereverberated, separated, and subjected to noise suppression using the techniques in Fig. 2. The adaptation was performed by using maximum likelihood linear regression with automatically obtained multiple regression matrices [10].

We employ two types of language models. One is a standard back-off n-gram model. As with acoustic modeling, it is difficult to obtain a meeting transcript that is large enough to estimate the n-gram model. We use several types of data sets including a large written-text corpus and a small meeting transcript, and combine them with different weights based on an Expectation-Maximization algorithm. The other is a discriminative language model (DLM) trained with the R2D2 criterion [11]. This criterion is effective for training a language model that directly reduces recognition errors in a baseline speech recognizer.

The decoder is based on efficient WFST-based one-pass decoding [8] in which fast on-the-fly composition can be used for combining WFSTs such as $HCLG_1$ and $G_{3/1}$ during decoding, where $HCLG_1$ represents a WFST that transduces an HMM state sequence into a unigram-weighted word sequence, and $G_{3/1}$ represents a WFST that weights a word sequence with the trigram probabilities divided by the unigram probabilities. This division is necessary to cancel out the unigram probabilities already contained in $HCLG_1$.

Since the algorithm can handle any number of WFSTs for composition on the fly, we combine four WFSTs, including two WFSTs that represent a DLM in the one-pass decoding. The first DLM WFST is based on word features and the second is based on part-of-speech (POS) features. Since a DLM can be represented as a set of word or POS n-grams with certain weights, it can be transformed into a WFST in the same way as a standard back-off n-gram model. The two DLM WFSTs for the word and POS features

can be combined linearly by the composition operation during decoding. Thus, we can achieve one-pass real-time speech recognition using these DLM WFSTs unlike the conventional approaches that involve a rescoring step after the first-pass decoding.

As part of the recent advances made in speech recognition technology in our research group, we have proposed an all-in-one speech recognition model [12], which is a different approach from those in which acoustic and language models are separately trained. The all-in-one model is represented as a WFST (or a set of WFSTs) including all the acoustic, pronunciation, and language models, and it is effectively trained using a discriminative criterion to reduce the number of recognition errors. As a result, the model can capture not only the general characteristics of acoustic and linguistic patterns but also a wide variety of interdependencies between acoustic and linguistic patterns in conversational speech. With this model, we have improved the speech recognition accuracy for meeting speech, and have increased the accuracy substantially by integrating the model with deep learning techniques [13].

2.4 Diarization-based word filtering

Finally, we describe diarization-based word filtering (DWF), which is an important technique used with our meeting speech recognizer. As shown in Fig. 2, we employed speech recognition for each speaker's channel given by source separation. This approach is much more robust in the case of overlapping speech than that using only a single channel. However, even if we employ source separation, it is difficult to completely remove other speakers' voices from the target speaker's channel. Such remaining non-target speech signals often induce insertion errors in speech recognition. To solve this problem, we utilize frame-based speaker diarization results obtained using the method shown in Fig. 2 to reduce the number of insertion errors. Since the diarization result at each frame tends to be discontinuous on the time axis, we use the average value of the diarization results for each recognized word, which can be considered to represent the relevance of the word to the target speaker. With this measure, words with a low relevance can be effectively deleted from the recognition results.

The relevance of word w_n recognized for speaker n is computed as:

$$s(w_n) = \frac{1}{e(w_n) - b(w_n) + 1} \sum_{t=b(w_n)}^{e(w_n)} d_n(t)$$

where $d_n(t)$ is the frame-based diarization result at frame t, and $b(w_n)$ and $e(w_n)$ respectively indicate the beginning and ending frames of word w_n. If $s(w_n)$ is less than a predefined threshold, w_n is deleted from the speech recognition result. This method is effective for low-latency processing. Conventional methods detect a speech segment by speaker diarization, and then apply speech recognition for the segment. This requires a long time delay because speech recognition cannot start until the diarization step is finished. The DWF approach only requires a one-frame (32 msec) delay to obtain $d_n(t)$, and it can drive speech recognition in parallel.

3. Real-time meeting analysis system

Our meeting analyzer basically recognizes *who is speaking what* by using speech recognition and speaker diarization, and it detects the activity of each participant (e.g., speaking, laughing, looking at someone else) and the circumstances of the meeting (e.g., topic, activeness, casualness, and intelligibility (defined more specifically below)) by integrating results obtained from several processing modules. The detected results provide *speaking to whom and how* information. The results of analysis are continuously displayed on a browser running on an Android™* tablet, as shown in **Fig. 3**.

The panel on the left side of the browser displays live streaming video of a 360-degree view provided by a camera together with information about *who is speaking to whom* represented by orange circles and light blue arrows. The right side shows the real-time transcript for each participant, as well as his/her picture. The face icon beside the transcript indicates the participant's state (speaking, laughing, or silent), and the next two bars show the number of words spoken by the participant and how much he/she has been watched by others, i.e., the visual focus of attention for each speaker, based on the direction of each participant's face.

The lower left panel shows the current circumstances of the meeting in terms of topic words and the degrees of activeness, casualness, and intelligibility. Activeness is calculated as the number of words spoken multiplied by the entropy based on the relative frequencies of spoken words for all the speakers in a fixed time window. Casualness is estimated based on the frequency of laughter. Intelligibility is calculated based on the frequency of participants' nods. These

* Android™ is a trademark of Google.

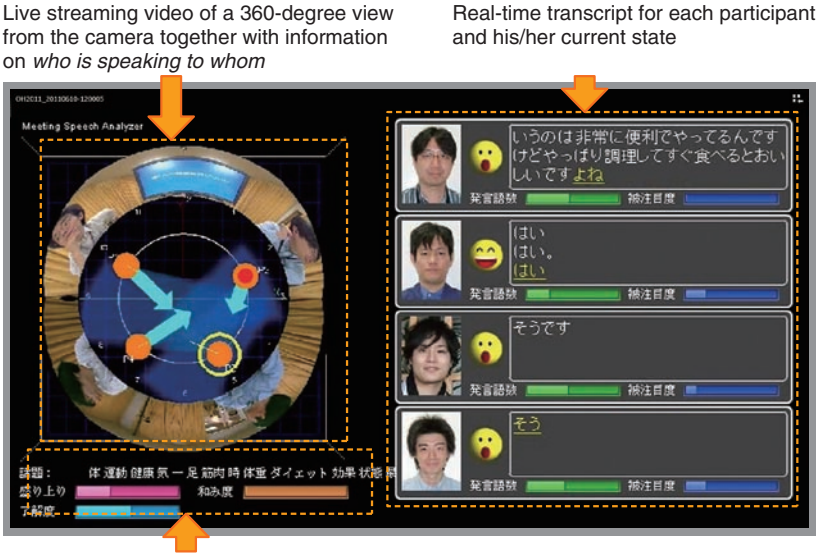


Fig. 3. Real-time meeting browser working with Android™.

graphical representations help us to understand the current circumstances of the meeting both visually and objectively.

The system architecture is depicted in **Fig. 4**. In the speech enhancement block, the m-th microphone signal in the STFT domain $y_m(f, t)$ ($m = 1, \cdots, 8$) is dereverberated, separated, and denoised, and finally the enhanced signals $\hat{s}_n(\tau)$ are sent to the speech recognizer and the acoustic event detector. Here, f and t are frequency and time-frame indices, respectively. Speech recognition is used for each separated signal to transcribe utterances. The speech recognizer SOLON is also used to detect acoustic events including silence, speech, and laughter. Sentence boundary detection and topic tracking are applied to the word sequences from SOLON. In topic tracking, we use the Topic Tracking Language Model [14], which is an online extension of latent Dirichlet allocation that can adaptively track changes in topics by considering the information history of the meeting. In our system, we use conditional random fields for sentence boundary detection [15], where the transition features consist of bigrams of the labels that identify the presence or not of a sentence head. The other features consist of words and their POS tags in the scope of a 3-word context and the pause duration at each boundary candidate.

The camera captures a 360-degree view from the center of the table. During visual processing, the faces of the participants are detected, and face images are sent to the browser at the beginning of the meeting. Then the face pose tracker continues to work during the meeting. This incorporates a Sparse Template Condensation Tracker [1], which realizes the real-time robust tracking of multiple faces by utilizing GPUs (graphics processing units). With this tracking approach, the position of each participant and his/her face direction can be obtained continuously. This visual information is used to determine *who is speaking to whom* and to detect the visual focus of attention. The position information is also used to associate each utterance with the corresponding participant by combining it with the DOA information of the speech signal. Accordingly, a transcript and a face icon can be displayed on the right panel for the participant who is speaking.

The meeting analysis module calculates the activeness and casualness of the meeting based on speech recognition and acoustic event detection results. The intelligibility is obtained based on the frequency of participants' nods detected by face-pose tracking. All the analysis results are sent to the browser together with streaming video via a real-time messaging protocol (RTMP) server. Since the RTMP server can receive multiple requests, the analysis results can be broadcast to multiple browsers.

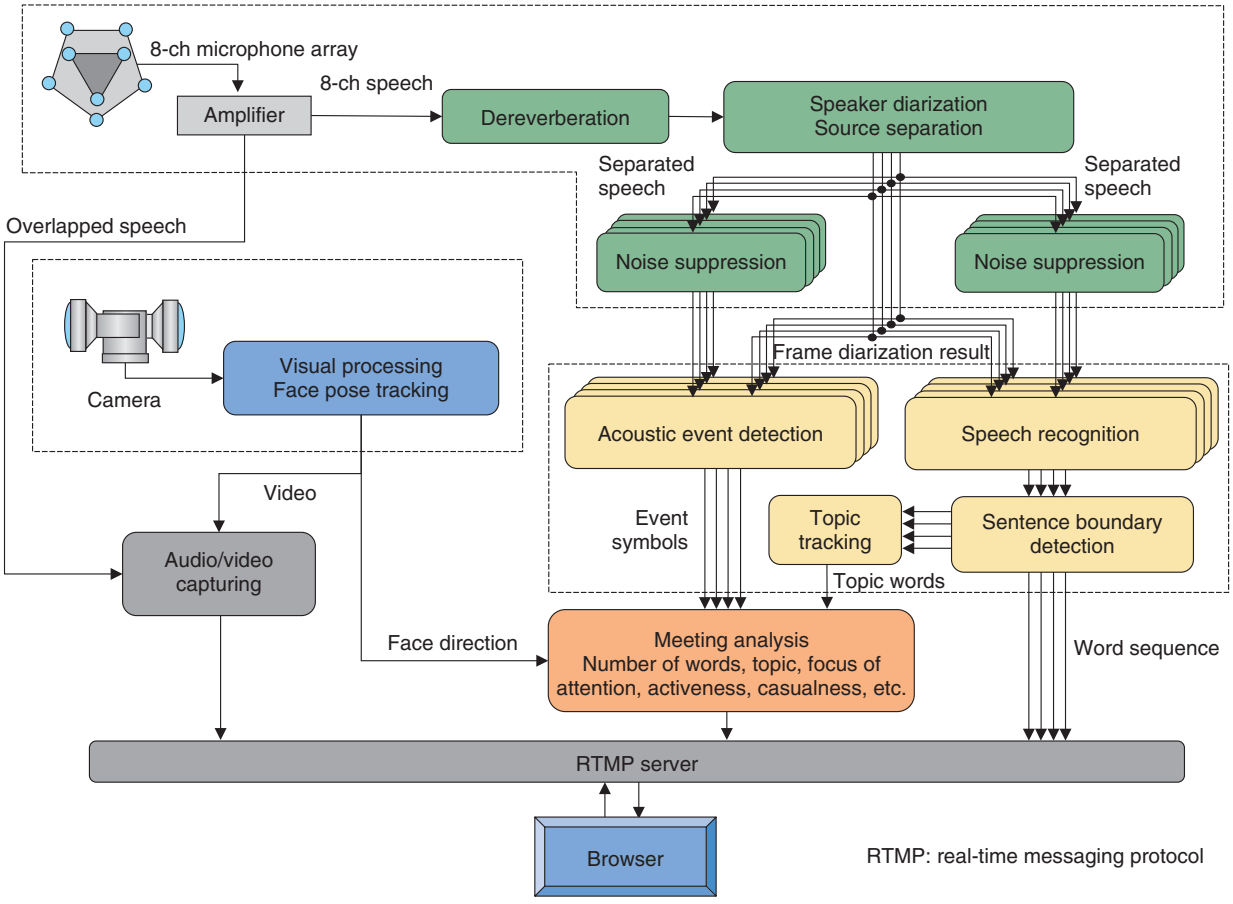


Fig. 4. Architecture of real-time meeting analysis system.

The current system runs on four computers: (1) an AMD Opteron 1389 2.9-GHz Quad Core for speaker diarization and source separation, (2) an Intel Xeon X5570 2.93 GHz 8-core dual processor for dereverberation and noise suppression, (3) the same Xeon model for speech recognition, acoustic event detection, and meeting analysis, and (4) an Intel Core 2 Extreme QX9650 3.0-GHz (+ NVIDIA GeForce-9800GX2/2 GPU cores) for visual processing.

4. Conclusion

In this article, we reviewed the technology we have developed in our research group and presented our meeting analysis system that provides accurate recognition of *who spoke when, what, to whom, and how* in a meeting situation. If conversational speech recognition and understanding based on audio-visual scene analysis becomes possible, many useful applications could be realized. In the future, it may be possible not only to generate meeting minutes automatically, but also to easily find past meeting scenes when required. We might have a virtual secretary who could answer our questions and register our plans in the scheduler autonomously. To realize such a system, it is important to improve speech recognition accuracy, and also to detect what is occurring in the surroundings, how the speaker is feeling, and why the meeting led us to a certain conclusion, etc. To enable such a deep understanding of human conversation, we need to extend the meeting analysis technology so that it can recognize higher-level concepts. To this end, we are continuing to address problems beyond the framework of speech recognition technology.

References

[1] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, "A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization," Proc. of the 10th International Conference on Multimodal Interfaces (ICMI 2008),

pp. 257–264, Chania, Greece.
[2] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 2, pp. 499–513, 2012.
[3] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 24, No. 4, pp. 320–327, 1976.
[4] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "Speaker indexing and speech enhancement in real meeting/conversations," Proc. of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), Vol. 1, pp. 93–96, Las Vegas, NV, USA.
[5] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A study of mutual frontend processing method based on statistical model for noise robust speech recognition," Proc. of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009), pp. 1235–1238, Brighton, UK.
[6] 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
[7] T. Nakatani, T. Yoshioka, S. Araki, M. Delcroix, and M. Fujimoto, "LogMax observation model with MFCC-based spectral prior for reduction of highly nonstationary ambient noise," Proc. of the 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), pp. 4029–4032, Kyoto, Japan.
[8] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 15, No. 4, pp. 1352–1365, 2007.
[9] E. McDermott, S. Watanabe, and A. Nakamura, "Discriminative training based on an integrated view of MPE and MMI in margin and error space," Proc. of the 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010), pp. 4894–4897, Dallas, TX, USA.
[10] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, Vol. 9, pp. 171–185, 1995.
[11] T. Oba, T. Hori, A. Nakamura, and A. Ito, "Round-robin duel discriminative language models," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 4, pp. 1244–1255, 2012.
[12] Y. Kubo, S. Watanabe, T. Hori, and A. Nakamura, "Structural classification methods based on weighted finite-state transducers for automatic speech recognition," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 8, pp. 2240–2251, 2012.
[13] Y. Kubo, T. Hori, and A. Nakamura, "Large vocabulary continuous speech recognition based on WFST structured classifiers and deep bottleneck features," Proc. of the 38th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), pp. 7629–7632, Vancouver, Canada.
[14] S. Watanabe, T. Iwata, T. Hori, A. Sako, and Y. Ariki, "Topic tracking language model for speech recognition," Computer Speech and Language, Vol. 25, No. 2, pp. 440–461, 2011.
[15] T. Oba, T. Hori, and A. Nakamura, "Sentence boundary detection using sequential dependency analysis combined with CRF-based chunking," Proc. of the 9th International Conference on Spoken Language Processing (INTERSPEECH 2006), pp. 1153–1156, Pittsburgh, PA, USA.

Takaaki Hori
Senior Research Scientist, Signal Processing Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E. and M.E. degrees in electrical and information engineering and the Ph.D. degree in system and information engineering from Yamagata University in 1994, 1996, and 1999, respectively. He joined NTT in 1999 and began researching spoken language processing at NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories). He moved to NTT Communication Science Laboratories in 2002. He was a visiting scientist at the Massachusetts Institute of Technology, Cambridge, MA, USA, from 2006 to 2007. He received the 22nd Awaya Prize Young Researcher Award from the Acoustical Society of Japan (ASJ) in 2005, the 24th TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2009, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan in 2012, and the 58th Maejima Hisoka Award from the Tsushinbunka Association in 2013. He is a member of ASJ, the Institute of Electronics, Information and Communication Engineers (IEICE), and IEEE.

Shoko Araki
Senior Research Scientist, Signal Processing Group, Media Information Laboratory, NTT Communication Science Laboratories.

She received the B.E. and M.E. degrees from the University of Tokyo in 1998 and 2000, respectively, and the Ph.D. degree from Hokkaido University in 2007. Since joining NTT in 2000, she has been engaged in research on acoustic signal processing, array signal processing, blind source separation (BSS) applied to speech signals, meeting diarization, and auditory scene analysis. She has been a member or had chairing roles in various committees and conferences, including ICA 2003, IWAENC 2003, EUSIPCO 2006, WASPAA 2007, and SiSEC 2008, 2010, and 2011. She received the 19th Awaya Prize Young Researcher Award from ASJ in 2001, the Best Paper Award of IWAENC in 2003, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004, the Academic Encouraging Prize from IEICE in 2006, and the Itakura Prize Innovative Young Researcher Award from ASJ in 2008. She is a member of ASJ, IEICE, and IEEE.

Tomohiro Nakatani
Senior Research Scientist, Supervisor, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E., M.E., and Ph.D. degrees from Kyoto University in 1989, 1991, and 2002, respectively. Since joining NTT as a researcher in 1991, he has been investigating speech enhancement technologies for developing intelligent human-machine interfaces. During 2005–2006, he was a Visiting Scholar at the Georgia Institute of Technology, Atlanta, GA, USA. Since 2008, he has been a Visiting Assistant Professor in the Department of Media Science, Nagoya University. He received the 1997 the Japanese Society for Artificial Intelligence Conference Best Paper Award, the 2002 ASJ Poster Award, the 2005 IEICE Best Paper Award, and the 2009 ASJ Technical Development Award. He has been a member of IEEE SP Society Audio and Acoustics Technical Committee (AASP-TC) since 2009 and a chair of the AASP-TC Review Subcommittee since 2013. He served as an Associate Editor of the IEEE Transactions on Audio, Speech, and Language Processing during 2008–2011 and has chaired or co-chaired several committees and conferences, including the IEEE Kansai Section Technical Program Committee, IEEE WASPAA-2007, and the IEEE CAS Society Blind Signal Processing Technical Committee. He is a senior member of IEEE, and a member of ASJ and IEICE.

Atsushi Nakamura
Senior Research Scientist, Supervisor, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E., M.E., and Dr.Eng. degrees from Kyushu University, Fukuoka, in 1985, 1987, and 2001, respectively. In 1987, he joined NTT, where he engaged in R&D of network service platforms, including studies on application of speech processing technologies to network services at Musashino Electrical Communication Laboratories. From 1994 to 2000, he was a Senior Researcher at Advanced Telecommunications Research (ATR) Institute, Kyoto, Japan, where he was engaged in spontaneous speech recognition research, construction of a spoken language database, and development of speech translation systems. Since April 2000, he has been with NTT Communication Science Laboratories. His research interests include acoustic modeling of speech, speech recognition and synthesis, spoken language processing systems, speech production and perception, computational phonetics and phonology, and application of learning theories to signal analysis and modeling. He received the IEICE Paper Award in 2004, and twice received the Telecom-technology Award of The Telecommunications Advancement Foundation, in 2006 and 2009. He is a senior member of IEEE and serves as a member of the IEEE Machine Learning for Signal Processing (MLSP) Technical Committee, and as the Chair of the IEEE Signal Processing Society Kansai Chapter. He is also a member of ASJ and IEICE.

Speech Recognition Based on Unified Model of Acoustic and Language Aspects of Speech

Yotaro Kubo, Atsunori Ogawa, Takaaki Hori, and Atsushi Nakamura

Abstract

Automatic speech recognition has been attracting a lot of attention recently and is considered an important technique to achieve natural interaction between humans and machines. However, recognizing spontaneous speech is still considered to be difficult owing to the wide variety of patterns in spontaneous speech. We have been researching ways to overcome this problem and have developed a method to express both the acoustic and linguistic aspects of speech recognizers in a unified representation by integrating powerful frameworks of deep learning and a weighted finite-state transducer. We evaluated the proposed method in an experiment to recognize a lecture speech dataset, which is considered as a spontaneous speech dataset, and confirmed that the proposed method is promising for recognizing spontaneous speech.

Keywords: speech recognition, deep learning, spontaneous speech

1. Introduction

Automatic speech recognition refers to the technology that enables a computer to extract and identify the words contained in given speech signals. Recently, systems using speech recognition technology have become increasingly common and have been used in several real world applications.

In addition to the development of systems and applications, basic techniques to enable accurate recognition have also been intensively investigated. In the first era of speech recognition technology, speech recognizers were only able to recognize speech uttered by one preregistered person. In the last decade, however, speech recognizers have become more powerful, and this has enabled recognition of speech from unknown persons if the input speech signals are appropriately uttered. Consequently, the current state-of-the-art technologies focus mainly on recognition of *spontaneous speech.*

Recognition of spontaneous speech is difficult because the assumption we introduced above, that speech is *appropriately uttered*, no longer holds in this case. The main objective of research on spontaneous speech recognition is to recognize speech signals even though they are inappropriately uttered but are still possible to be perceived by humans. Such inappropriate speech is characterized by several fluctuations in the input signals. For example, utterances such as *ah, well, uh, or hmm*, which are called fillers, are frequently inserted, and sometimes several articles and/or particles are deleted. Since speech recognizers recognize speech signals in accordance with linguistic rules, these fluctuations lead to recognition errors. Furthermore, fluctuations in pronunciation also affect speech recognizers. For example, even though a human may perceive that speech signals have been pronounced correctly, the computer analysis results often fluctuate for several reasons such as vowel omission or unstable vocal tract control. These fluctuations in acoustic and linguistic aspects of speech make recognition of spontaneous speech

difficult.

Conventionally, linguistic fluctuations of human speech are expressed by using probabilistic models called *language models*, and acoustic fluctuations are expressed separately by *acoustic models.*

However, as mentioned above, because the fluctuations that occur with spontaneous speech often appear in both the acoustic and linguistic aspects, the recognition of spontaneous speech by using such probabilistic models based on a divide-and-conquer strategy is considered to be difficult.

Deep learning techniques, which integrate signal processing models and acoustic models, have recently demonstrated significant improvement over conventional speech recognizers that have separate signal processing and acoustic models [1]. Deep learning suggests that optimizing several models in a unified way is important in order to overcome such difficult phenomena in spontaneous speech recognition. However, even with these advanced techniques, the fluctuations that span both the acoustic and linguistic aspects of speech have not yet been sufficiently expressed since deep learning techniques do not optimize linguistic aspects of recognizers.

In this article, we describe a method that enables joint optimization of the acoustic and language models of speech recognition, and we explain how this technique improves speech recognizers for spontaneous speech by focusing on speech recognition of a lecture video.

2. Weighted finite-state transducers

Weighted finite-state transducers (WFSTs) are commonly used as a core software component of several automatic speech recognizers including our proposed speech recognizers. WFSTs are abstract machines that represent rules to convert one type of sequence into another type of sequence, for example, to convert a speech feature sequence into a word sequence in automatic speech recognition. All the probabilistic models used in automatic speech recognizers can be expressed by using WFSTs, and therefore, WFSTs are used as a unified representation of automatic speech recognizers [2]. The speech recognition technique we describe in this article enabled joint acoustic and linguistic representation by extending WFSTs.

An illustration of an example WFST that converts phoneme sequences to word sequences is shown in **Fig. 1(a)**. The circles in the figure represent internal states of the abstract machine, and arrows indicate that the state may change in the direction of the arrow. The numerical values annotated to the arrows denote probabilities of the state transition corresponding to the arrow, and the symbols annotated to the arrow (for example, "ow"/"go") mean that the machine is expected to read the symbol "ow" from the input sequence during this state change, and to write the symbol "go" to the output sequence. The conversion process starts from the initial state (state 1), follows the arrow repeatedly while reading from the input sequence and writing to the output sequence, and ends when the state reaches the final state (state 7).

One of the main advantages of using WFSTs is that they have advanced composition algorithms. A WFST that accepts the word sequences that can be assumed as system inputs is shown in **Fig. 1(b)**. Even though this WFST actually does no conversion (i.e., it only outputs the same sequence as the input), this kind of probabilistic acceptance can also be represented in a WFST. The composition algorithm processes these two WFSTs (Figs. 1(a) and (b)) and constructs a composite WFST as in **Fig. 1(c)**. The composite WFST is constructed to represent the cascade connection of the input WFSTs. In other words, the WFST represents all possible conversion patterns obtained if the output sequences of the WFST in Fig. 1(a) are used as input sequences of the WFST in Fig. 1(b). The probability corresponding to each conversion pattern is simply denoted as a product of these two internal transductions.

The entire probabilistic process of automatic speech recognition can be represented by a large WFST that converts acoustic pattern sequences representing a short-time spectral pattern of acoustic signals to word sequences. This large WFST can be obtained by applying the composition algorithm to elemental WFSTs that convert acoustic patterns to interim representations called phoneme-states, the phoneme-states to phonemes, the phonemes to words (Fig. 1(a)), and the words to word sequences (Fig. 1(b)), respectively. Automatic speech recognition is subsequently achieved by finding the path on the WFST that maximizes that probability.

Estimating the elemental WFSTs (Figs. 1(a) and (b) in this case) is the central problem in constructing speech recognizers. Typically, these WFSTs are constructed by converting probabilistic models corresponding to each element into a WFST representation.

However, the strategy based on combining elemental WFSTs that are estimated separately cannot sufficiently express the phenomena in spontaneous

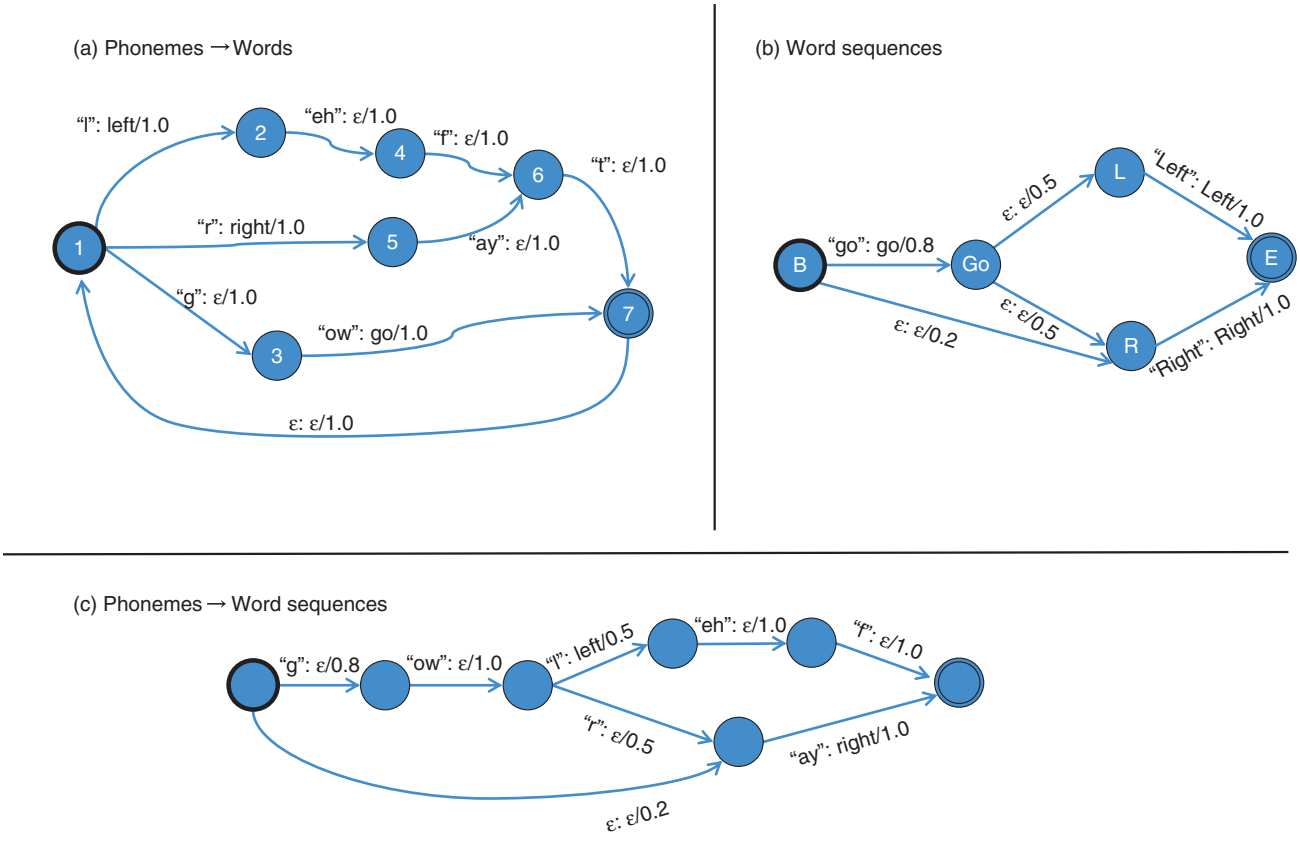


Fig. 1. Example of WFSTs.

speech because these phenomena often span several elemental WFSTs. Thus, it is necessary to consider the interdependency of these elemental WFSTs in order to model these fluctuations.

3. Acoustic model based on deep learning

Recently, researchers at Toronto University applied a method called *deep learning* to achieve accurate acoustic modeling. It has been demonstrated that deep learning methods can achieve accurate speech recognition without the need for complex acoustic pattern normalization techniques. Deep learning is a general term that refers to advances in the study of neural networks that have relatively deep architectures. Even though research on neural networks has been going on for over 30 years, the practical use of models with deep architectures was impossible before deep learning was developed.

The neural networks we focused on compute the output vector $y = (y_1, y_2, \ldots, y_{D'})^T$ with the given input vector $x = (x_1, x_2, \ldots, x_D)^T$ by using the following equation

$$y_j(x) = h_j^{(L)}(x),$$

$$h_j^{(\ell)}(x) = f\left(\sum_{i=1}^{D^{(\ell)}} w_{i,j}^{(\ell)}\ h_j^{(\ell-1)}(x) + b_j^{(\ell)}\right),$$

$$h_j^{(0)}(x) = x_i,$$

$$f(z) = \frac{1}{1 + e^{-z}},$$

where e is Napier's constant (the base of the natural logarithm).

By optimizing the parameters of the above equation ($w_{i,j}^{(\ell)}$ and $b_j^{(\ell)}$) so that the equation represents the given examples of x and y, we can use this equation to predict y that corresponds to the unseen example x. In automatic speech recognizers, x typically denotes the vectors that represent speech signals, and y typically denotes the probability of appearance of each acoustic pattern. The above equation includes L recursions, and L should be adjusted manually.

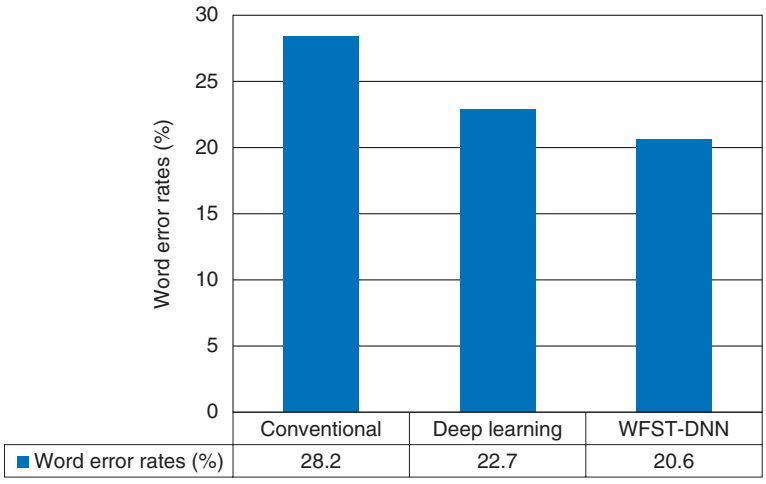


Fig. 2. Word error rates of three methods in English lecture speech recognition experiment.

Deep learning techniques enable optimization of neural networks with large L (such as $L > 3$) that are conventionally considered difficult to optimize. To enable such optimization, deep learning introduces an additional training procedure called pretraining. In this procedure, the parameters of neural networks are optimized so that the input vectors x in training examples are accurately expressed in the neural networks. Performing this pretraining procedure before the actual optimization procedure, which optimizes networks so that correspondence of x and y is expressed in the networks, makes it possible to optimize neural networks with large L, which are called deep neural networks (DNNs).

Since DNNs can be viewed as composite models of pattern correctors, the optimization of DNNs can be viewed as a method to achieve unified modeling of pattern normalization and classification.

4. Unified model based on deep learning

An entire conversion process from acoustic patterns to word sequences can be represented in a unified form by using large WFSTs. Furthermore, recent advances in deep learning have led to the development of a unified algorithm for acoustic pattern normalization and acoustic pattern classification. To take advantage of both approaches, we developed a unified modeling technique called WFST-DNN that integrates pattern normalization, pattern classification, acoustic models, and language models defined by unifying WFST and deep learning technologies [5].

In WFST-DNN, we enhanced the probabilities annotated to each arc of the WFSTs. In conventional methods, these probabilities are computed as a product of the probabilities of each elemental WFST. We enhanced these probabilities by defining them as an output of DNNs. Specifically, we defined and optimized y in the above equation to represent the probability of arcs. By applying this enhancement, the implicit assumption introduced in the conventional method, which is that the fluctuations appear independently in each elemental WFST, can be prevented. This enhancement is straightforward in that the speech fluctuations caused by phenomena of spontaneous speech span both the acoustic and language models. Further, since the proposed approach does not change the structure of conventional WFSTs, advanced methods, for example, computationally efficient recognition techniques developed for WFST-based speech recognizers can also be applied to a speech recognizer with the proposed technique.

We applied this method to a lecture recognition task and found that it performed better than a conventional method. The word error rates of the proposed method and the conventional method are shown in **Fig. 2.** Here, *conventional method* denotes a conventional state-of-the-art method before the introduction of deep learning techniques [6], and *deep learning* is of course the method based on deep learning. The word error rate achieved using the deep learning method was surprisingly high. WFST-DNN denotes the proposed method. It is clear from the results that the proposed method exhibited improved performance compared to the advanced conventional system and the results of the deep learning method.

5. Future outlook

The two main objectives of our future research are as follows. The first one is to achieve a deeper understanding of deep learning techniques. Deep learning techniques involve difficulties in terms of mathematical analysis, and therefore, the advantages of deep learning have only been shown through empirical and experimental results. However, investigating the advantages of deep learning and understanding how deep learning achieves advanced acoustic modeling would be very useful in order to apply these techniques in many other fields.

The second objective is to develop more computationally efficient techniques. Even though the current WFST-DNN recognizer can output recognition results in an acceptable time frame by exploiting graphic processing units (GPUs), it is important to be able to compute the results in a personal computer without fast GPUs. Parameter optimization requires a very long time even with acceleration based on GPUs. This long optimization time would also be problematic when customizing the system for a specific application.

We will pursue these objectives as we continue to investigate speech recognition techniques, with the ultimate goal of enabling application to various fields and achieving accurate recognition.

References

[1] F. Seide, G. Li, and D. Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," Proc. of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Florence, Italy.

[2] T. Hori and A. Nakamura, "Speech Recognition Algorithms Using Weighted Finite-State Transducers," Morgan & Claypool Publishers, 2013.

[3] A.-R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 1, pp. 14–22, 2012.

[4] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.

[5] Y. Kubo, T. Hori, and A. Nakamura, "Integrating Deep Neural Networks into Structured Classification Approach Based on Weighted Finite-State Transducers," Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012), Portland, OR, USA.

[6] E. McDermott, S. Watanabe, and A. Nakamura, "Discriminative Training Based on an Integrated View of MPE and MMI in Margin and Error Space," Proc. of 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 4894–4897, Dallas, TX, USA.

Yotaro Kubo
Research Engineer, Media Information Processing Laboratory, NTT Communication Science Laboratories.

He received the B.E., M.E., and Dr.Eng. degrees from Waseda University, Tokyo, in 2007, 2008, and 2010, respectively. He was a visiting scientist at RWTH Aachen University, Aachen, Germany, from April to October 2010. In 2010, he joined NTT and has been with NTT Communication Science Laboratories. His research interests include machine learning and signal processing. He received the Awaya Award and the Itakura Prize Innovative Young Researcher Award from the Acoustical Society of Japan (ASJ) in 2010 and 2013, respectively, the Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) in 2011, and the IEEE Signal Processing Society Japan Chapter Student Paper Award in 2011. He is a member of the International Speech Communication Association, ASJ, IPSJ, the Institute of Electronics, Information and Communication Engineers (IEICE), and IEEE.

Atsunori Ogawa
Research Engineer, Media Information Processing Laboratory, NTT Communication Science Laboratories.

He received the B.E. and M.E. degrees in information engineering, and the Ph.D. degree in information science from Nagoya University, Aichi, in 1996, 1998, and 2008, respectively. Since joining NTT in 1998, he has been engaged in research on speech recognition. He received the ASJ Best Poster Presentation Award in 2003 and 2006, respectively. He is a member of ASJ, IPSJ, IEICE, and IEEE.

Takaaki Hori
Senior Research Scientist, Signal Processing Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E. and M.E. degrees in electrical and information engineering and the Ph.D. degree in system and information engineering from Yamagata University in 1994, 1996, and 1999, respectively. He joined NTT in 1999 and began researching spoken language processing at NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories). He moved to NTT Communication Science Laboratories in 2002. He was a visiting scientist at the Massachusetts Institute of Technology, Cambridge, MA, USA, from 2006 to 2007. He received the 22nd Awaya Prize Young Researcher Award from ASJ in 2005, the 24th TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2009, the Kiyasu Special Industrial Achievement Award from IPSJ in 2012, and the 58th Maejima Hisoka Award from the Tsushinbunka Association in 2013. He is a member of ASJ, IEICE, and IEEE.

Atsushi Nakamura
Senior Research Scientist, Supervisor, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E., M.E., and Dr.Eng. degrees from Kyushu University, Fukuoka, in 1985, 1987, and 2001, respectively. In 1987, he joined NTT, where he engaged in R&D of network service platforms, including studies on application of speech processing technologies to network services at Musashino Electrical Communication Laboratories. From 1994 to 2000, he was a Senior Researcher at Advanced Telecommunications Research (ATR) Institute, Kyoto, Japan, where he was engaged in spontaneous speech recognition research, construction of a spoken language database, and development of speech translation systems. Since April 2000, he has been with NTT Communication Science Laboratories. His research interests include acoustic modeling of speech, speech recognition and synthesis, spoken language processing systems, speech production and perception, computational phonetics and phonology, and application of learning theories to signal analysis and modeling. He received the IEICE Paper Award in 2004, and twice received the Telecom-technology Award of The Telecommunications Advancement Foundation, in 2006 and 2009. He is a senior member of IEEE and serves as a member of the IEEE Machine Learning for Signal Processing (MLSP) Technical Committee, and as the Chair of the IEEE Signal Processing Society Kansai Chapter. He is also a member of ASJ and IEICE.

Speaking Rhythm Extraction and Control by Non-negative Temporal Decomposition

Sadao Hiroya

Abstract

Speaking rhythm plays an important role in speech production and the perception of non-native languages. This article introduces a novel method for extracting and controlling speaking rhythm from speech signals using non-negative temporal decomposition.

Keywords: articulatory movements, non-negative temporal decomposition, speaking rhythm

1. Introduction

Speech communication using non-native languages is difficult for many people both in speaking and listening to speech. By way of example, most native Japanese speakers have difficulty understanding what native English speakers are saying and therefore cannot communicate well in English with them. There are two major differences between Japanese and English: pronunciation (e.g., the number of vowels and the /R-L/ contrast) and rhythm. Pronunciation is very important for communication in English using words and short sentences (e.g., "Coffee, please." and "Where is the toilet?"). For long sentences, on the other hand, rhythm is more important than pronunciation. However, most Japanese learners of English regard pronunciation as important, rather than rhythm. As a result, native Japanese speakers have trouble communicating in English using long sentences with native English speakers.

In this article, I introduce a novel method of automatically correcting the halting English rhythm of native Japanese speakers by approximating the natural rhythm of native English speakers (**Fig. 1**).

2. Speaking rhythm

Rhythm generally refers to a pattern in time. In linguistics, languages can be categorized into two rhythms: stress-timed rhythm (e.g., English) and syllable-timed rhythm (e.g., Japanese). Chen et al. explained this as follows: "Stress-timed rhythm is determined by stressed syllables, which occur at regular intervals of time, with an uneven and changing number of unstressed syllables between them. Syllable-timed rhythm is based on the total number of syllables since each syllable takes approximately the same amount of time," [1] (**Fig. 2**). A syllable-timed rhythm is thus simpler than a stress-timed rhythm.

Humans follow a rhythm in various situations: speaking, playing musical instruments, clapping hands, walking, etc. Therefore, the definition of rhythm is not limited to only the temporal structure of sounds. In this study, I define speaking rhythm as a temporal pattern of movements made by articulatory organs such as the lips, jaw, tongue, and velum (soft palate); that is, not as sounds, but as articulatory movements. Some readers might question whether the definition of rhythm should be used for articulatory movements since most articulatory organs are inside the mouth, where speech is produced. However, I measured articulatory movements using an electromagnetic articulography (EMA) system and magnetic resonance imaging (MRI) [2] (**Fig. 3**) and estimated articulatory movements from speech signals [3]. My previous study indicated that articulatory movements are suitable for defining speaking rhythm [4]. Specifically, I analyzed the articulatory parameters

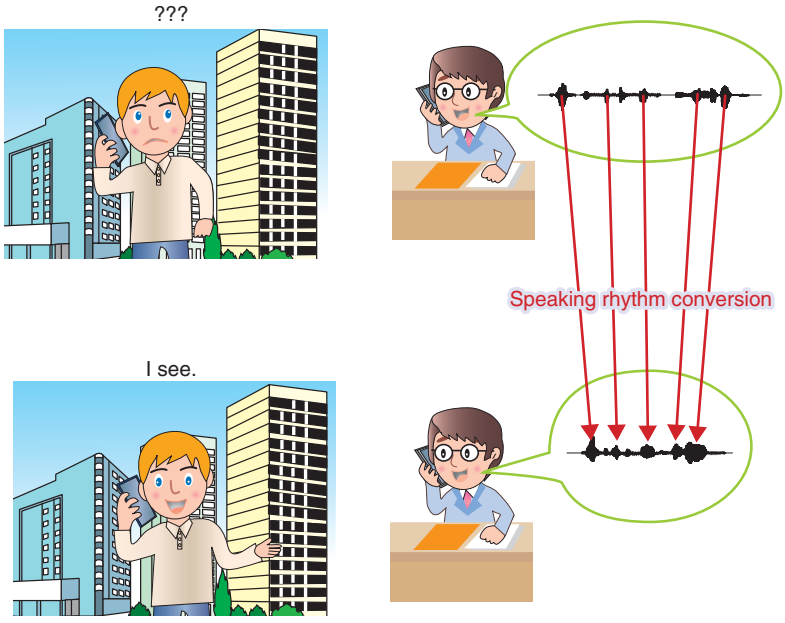


Fig. 1. Example of speaking rhythm conversion.

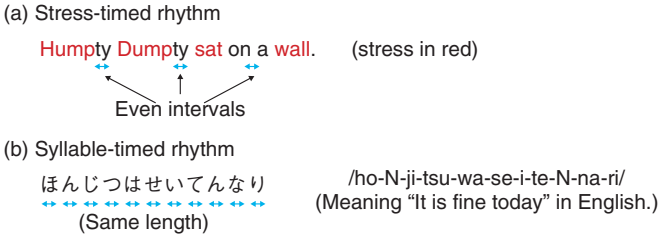
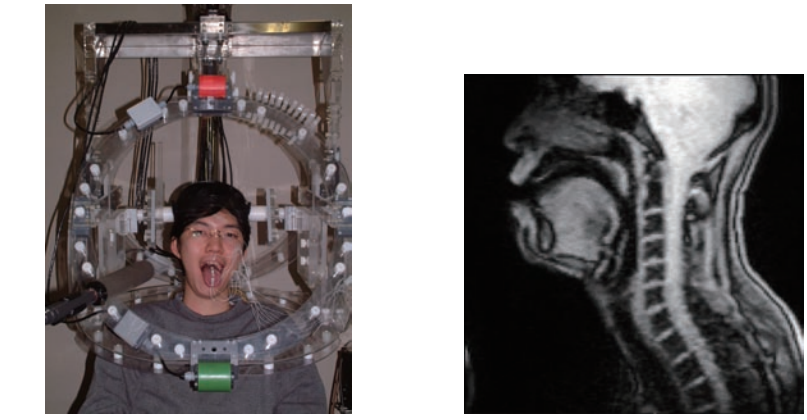


Fig. 2. Examples of stress-timed rhythm and syllable-timed rhythm.



(a) Electromagnetic articulography system (b) Magnetic resonance imaging*

* In collaboration with Konan University, Hyogo, Japan.

Fig. 3. Methods use to measure articulatory movements.

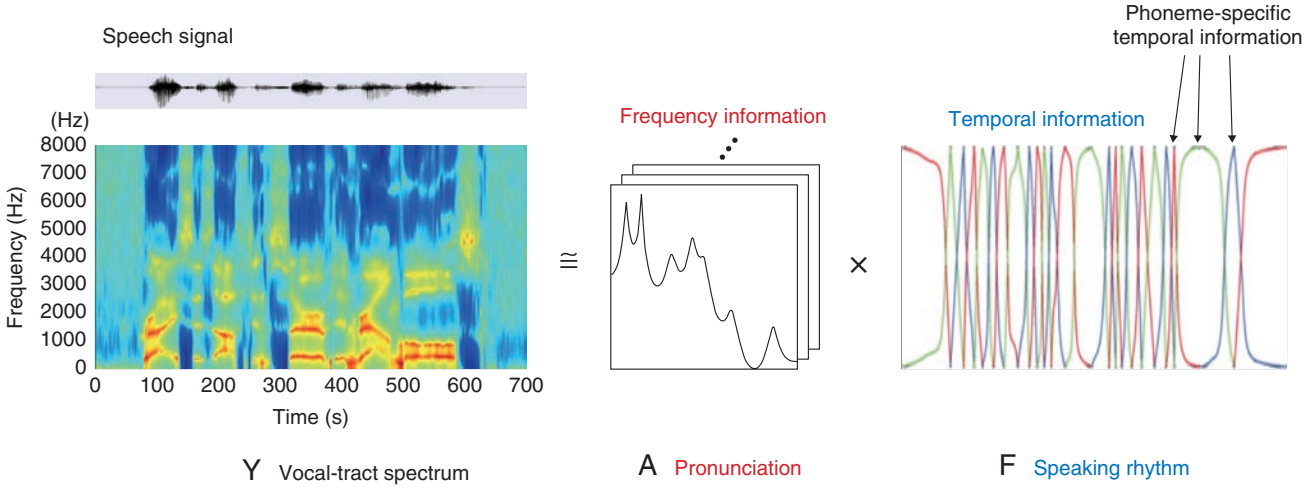


Fig. 4. Non-negative temporal decomposition.

represented by the vertical and horizontal positions of six receiver coils, which were placed on the lower incisor, the upper and lower lips, and the tongue (three positions), which were measured by the EMA system during speech production. The results revealed that articulatory parameters can be represented by articulatory positions at the central point of each phoneme and by linear interpolation. That is, a sparse representation of articulatory movements is suitable for obtaining speaking rhythm. This finding is related to articulatory phonology [5] and recent findings on the neural mechanism of speech production [6]. Also, this indicates that articulatory parameters change smoothly.

Consequently, treating articulatory movements as speaking rhythm should make it possible to easily extract and control speaking rhythm.

3. Non-negative temporal decomposition

Speech signals contain both frequency and temporal information. In audio signal processing, non-negative matrix factorization (NMF) can be applied to decompose audio signals into frequency and temporal information [7]. However, the NMF algorithm does not introduce articulatory-specific restrictions. Thus, it is not guaranteed that the temporal information will have a bell-shaped velocity profile, which is characteristic of human articulatory movements, and that only phonemes adjacent to the temporal information will affect it.

To overcome this problem, I developed a non-negative temporal decomposition (NTD) method to extract the speaking rhythm (temporal information) from speech signals under articulatory-specific restrictions.

NTD decomposes a vocal-tract spectrum (e.g., a line spectral pair), which is associated with articulatory organs, into a set of temporally overlapped phoneme-dependent event functions F and corresponding event vectors A under articulatory-specific restrictions (**Fig. 4**). Temporal information F introduces the phoneme-specific model and is affected only by adjacent phonemes. The NTD algorithm is as follows. First, a vocal-tract spectrum is calculated from speech signals. Then, an event function, which is restricted to the range [0,1], is determined by minimizing the squared Euclidean distance between the input and the estimated vocal-tract spectrum based on the multiplicative update rules in the NMF algorithm. Multiplicative update rules make it possible to obtain non-negative values of event functions and unimodal event functions without any penalty functions [4]. In fact, the multiplicative update rules would be more effective for improving the bell-shaped velocity profiles than a smoothing method with a penalty function introduced to NMF.

In NTD, the event timings need to be known. In this study, the timings were modified by minimizing the squared Euclidean distance by utilizing dynamic programming (DP). Thus, NTD can be considered a constrained NMF with DP. The only input for NTD is speech signals, but NTD can extract the speaking rhythm of articulatory movements due to the

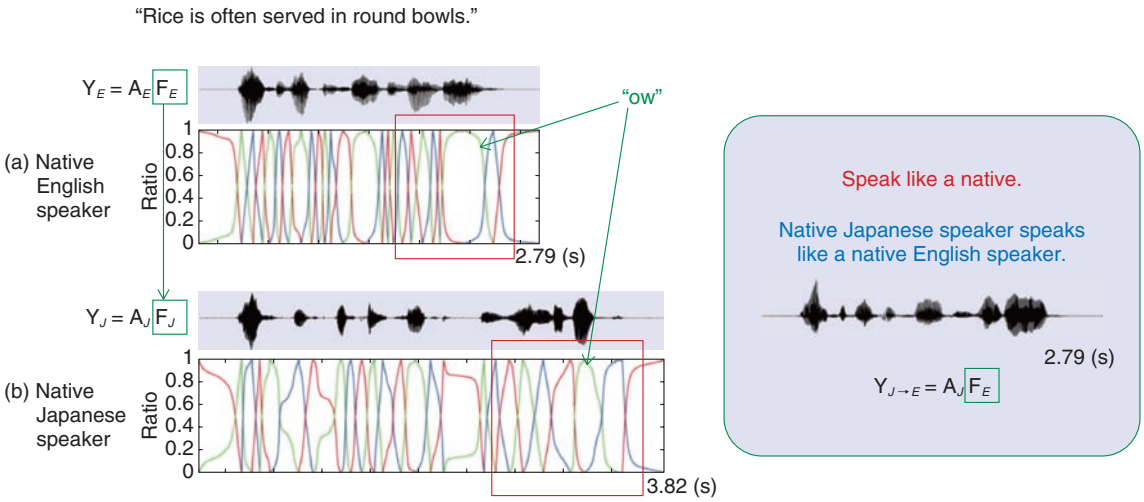


Fig. 5. Speaking rhythm conversion method.

articulatory-specific restrictions. Therefore, NTD is expected to be useful for acoustic-to-articulatory inversion [4].

4. Control of speaking rhythm

In this section, I explain how the speaking rhythm of an English sentence spoken by a native Japanese speaker is converted into the rhythm of a native English speaker (**Fig. 5**). First, both native Japanese and native English speakers read the same English sentence (e.g., "Rice is often served in round bowls".) Next, NTD is applied to extract frequency information A_J and temporal information F_J from the vocal-tract spectrum of the native Japanese speaker and to extract A_E and F_E from that of the native English speaker. I substitute F_E for F_J to obtain a vocal-tract spectrum with the pronunciation of native Japanese speaker A_J and the rhythm of native English speaker F_E. Finally, speech signals are generated from the vocal-tract spectrum and source signals. The generated speech signal in Fig. 5 appears to be time-compressed speech, in which the temporal characteristics of the speech signal are altered by reducing its duration without affecting the frequency characteristics. However, the temporal pattern in "bowls" (red square in Fig. 5) differs between the Japanese and English native speakers; the duration of "ow" for the English speaker is much longer than that for the Japanese speaker. This indicates that the technique is effective for controlling the English speaking rhythm of the native Japanese speaker. Feedback from native English speakers indicated that this speaking-rhythm-controlled speech signal using a personal computer was easier to understand.

5. Future prospects

Speech translation systems using another person's voice can also assist native Japanese speakers when they are communicating verbally in English. However, the opportunities for speech communication in English using one's own voice rather than another person's voice are expected to increase owing to the fact that English has been a required subject in elementary schools since 2011 in Japan.

This technique will be useful in practical applications such as communicating in English via teleconferences, public speaking, and using mobile phones. However, the technique cannot change the speaking rhythm of a sentence unless there is already a sample of the same sentence that has been read before. Thus, to become widely used, it will be necessary to model event functions between languages. I hope that this technique will eventually alleviate the burden involved in communication using non-native languages.

References

[1] C. Chen, C. Fan, and H. Lin, "A New Perspective on Teaching English Pronunciation: Rhythm," Proc. of the 4th International Symposium on English Teaching, pp. 24–41, Kaohsiung, Taiwan, 1996.

[2] S. Hiroya and T. Kitamura, "Generation of a vocal-tract MRI movie based on sparse sampling," Proc. of International Seminar on Speech Production (ISSP) 2011, pp. 1–8, Montreal, Canada.

[3] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," IEEE Trans. on Speech and Audio Processing, Vol. 12, No. 2, pp. 175–185, 2004.

[4] S. Hiroya, "Non-negative temporal decomposition of speech parameters by multiplicative update rules," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 21, No. 10, pp. 2108–2117, 2013.

[5] C. P. Browman and L. Goldstein, "Articulatory phonology: an overview," Phonetica, Vol. 49, No. 3-4, pp. 155–180, 1992.

[6] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation," Nature, Vol. 495, No. 7441, pp. 327–332, 2013.

[7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, Vol. 401, No. 6755, pp. 788–791, 1999.

Sadao Hiroya

Senior Research Scientist, NTT Communication Science Laboratories.

He received the B.S. degree from Tokyo University of Science in 1999 and the M.E. and Ph.D. degrees from Tokyo Institute of Technology in 2001 and 2006, respectively. He joined NTT Communication Science Laboratories in 2001. From 2001 to 2003, he was also a researcher in the CREST project of the Japan Science and Technology Agency. From 2007 to 2008, he was a visiting scholar at Boston University, MA, USA. In 2006, he received the 1st Itakura Prize Innovative Young Researcher Award from the Acoustical Society of Japan (ASJ). His current research interests include the links between production and perception of speech, functional brain imaging, and acoustic-to-articulatory inversion problems. He is a member of ASJ, the Institute of Electronics, Information and Communication Engineers, the Society for Neuroscience, and IEEE.

Link between Hearing and Bodily Sensations

Norimichi Kitagawa

Abstract

As humans, we know the size and shape of our own body, and we believe that our body is stable and maintains a consistent shape. However, some acoustic manipulations can induce illusions related to the body. These illusions indicate that hearing plays an important role in the sensations we perceive in our own body. This article presents an overview of such illusions and discusses the relationships between the sense of hearing and bodily sensations.

Keywords: hearing, bodily sensation, illusion

1. Introduction

Humans obtain information about the world through the so-called five senses (vision, hearing, touch, taste, and smell). Therefore, understanding the characteristics and mechanisms of these senses is essential for transmitting information from one person to another adequately. The role of the senses is widely recognized to be *to gain information regarding the world around us*. However, simply knowing our surrounding environment is not enough for us to act within the environment. We also must know about our own body. By knowing our own body, the surrounding environment, and the relationships between them, we are able to take appropriate action within the environment. We also obtain information about own body through our senses. This implies that by managing sensory information regarding the body, it is possible to control what a person feels (creating illusions) about his or her own body. It is well known that the somatic senses and vision play important roles in the perception of the body. In our research group, however, we have been conducting research focused on hearing. This article introduces the role that hearing plays in body sensations.

2. Bodily perception and vision

Of course, information about the body is obtained from the somatic senses, which relate to the body (such as the sense of touch, proprioception, and interoception). We can perceive something touching the body through the sense of touch, and the position and motion of different parts of the body through proprioception (the sense of extension of the skeletal muscles). Interoception gives information regarding the physiological state inside the body, and our sense of temperature tells us the temperature of our body and of objects that touch the skin. However, looking at the body is also an important way to know its state. For example, it is not possible to know the length of an arm or a body part accurately through somatic sensation alone. It has been shown that visual information is important in understanding body shape. Many visual clues are also used in determining whether something is touching the body.

The rubber hand illusion is a famous example of a phenomenon that shows the importance of vision in bodily perception [1]. The right hand of a participant is hidden, and a rubber model of a right hand is placed in front of him **(Fig. 1)**. The experimenter repeatedly strokes both the hand of the participant and the rubber hand in the same position at the same time. The participant sees the rubber hand being touched in the same way he feels his own hand being touched. After experiencing this for a few minutes, the subject starts to feel as though the sensation of being touched is coming from the rubber hand, and furthermore, that

the rubber hand is part of his own body. It has been reported that if the rubber hand is then subjected to an apparent *injury* while the illusion is in effect, the physiological responses to the injury are similar to those when the participants' own hand was injured [2]. When receiving both the tactile information from our own hand being touched and the visual information of the rubber hand being touched in the same way, our perceptual system interprets that the model hand is our own hand.

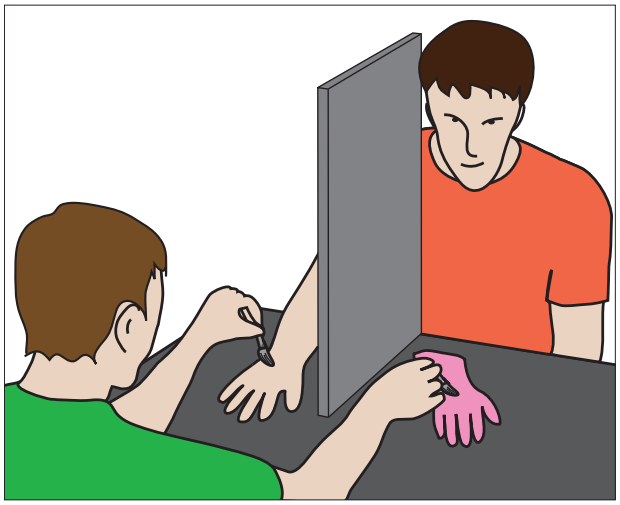


Fig. 1. The rubber hand illusion.

3. Touching with sound

To date, there has been very little research into the role played by the sense of hearing in how humans perceive the body. However, a sound is often produced when something touches the body or when we move our bodies, for example, the sound of footsteps when we walk. Such relationships between the body and sound suggest that the sense of hearing can contribute to bodily perception.

We conducted an experiment shown in **Fig. 2(a)** [3]. In the experiment, a microphone was placed in the ears of a dummy head, and the participating subjects were able to hear the sounds *heard* by the dummy ear. The dummy ear was then tickled using a small brush. The experiments were divided into two cases; in one case, the sound was presented very close to the subject's ear (i.e., through headphones), and in the other, the sound was presented from a loudspeaker located at a distance of 70 cm from the ear (with sound intensity equalized according to head position). Then, after listening to the sound for 30 seconds, the subjects were asked to rate several items on a scale of 1 to 7. The results are shown in **Fig. 2(b)**. When the sound was presented near the ear, the subjects felt as though their own ear was being tickled, but when it was presented at a distance, they did not perceive it that way. When they only viewed a video of the dummy ear being tickled but did not hear any sound, they did not feel as though they were being

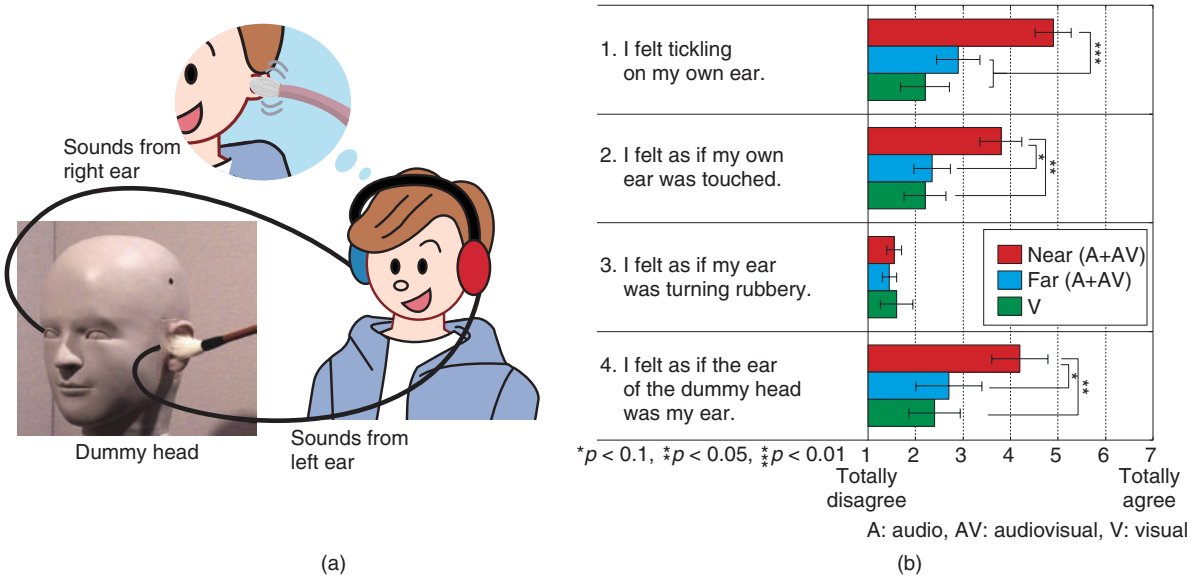


Fig. 2. Tactile illusion induced by sound. (a) Tickling with sound, (b) results of the experiment.

tickled. In the experiments in which the subjects only heard the sound and did not see the dummy being tickled, the subjects had no idea what the sound was. This shows that the subjects were not simply imagining they were being tickled when hearing the sound. When the sound was presented close to the ear, the subjects perceived that they were actually being touched, even though they were not. Various other tests were done that involved discriminating the locations and temporal order of tactile stimuli, and measuring simple response times for tactile stimuli and sounds. The results showed that when sound is presented close to the head, it affects the sense of touch in the vicinity of the sound [4]–[6].

The relationship between hearing and the sense of touch in the head is so strong that touch is perceived just by hearing a sound. However, it is also known that the sense of touch can change according to the sound when a sound is produced by a touch of another part of the body. For example, when the sound of rubbing both hands together is captured with a microphone and heard through headphones, if the high-frequency components are amplified, the listener's hands tend to feel dry, but if high-frequency components are attenuated, the hands are perceived as damp and heavy [7]. Sound effects in television and movies are well known, but sound effects can also affect the sense of touch in our hands. From the time we are born, whenever something touches our face or head, or when we touch something with our hands, and a sound is produced at the same time our skin feels a sensation, the sensations are experienced together, and we learn the relationship. Because of this relationship, a sensation of touch can be produced or changed by sound, even if we are not actually being touched.

4. Lengthening of arm through sound

Next, we introduce an illusion in which the perceived shape of the body is altered by hearing sounds. What we learned from the example of the rubber hand illusion introduced in Fig. 1 was that what we recognize as our own body is constantly being updated based on sensory information from somatic, visual, and other senses. If there are inconsistencies in the sensory information input, our awareness of *what our own body is* can change very easily. Analyzing illusions related to the body such as the rubber hand illusion reveals how we perceive our bodies.

Traditionally, the somatic senses such as touch and proprioception, as well as vision have been understood to be important in recognition of the body. However, we have recently discovered that hearing also contributes to recognition of the body [8]. We created conditions in which the sound of tapping on the floor is heard from a farther distance than where the tapping is actually occurring, as shown in **Fig. 3**. In the experiments, subjects were blindfolded and asked to kneel on the floor. An object was used to lightly touch the subjects' right forearms and left upper arms at two points ranging from 2–6 cm apart. This tactile stimulation was carried out both before and after the subjects listened to a sound coming from speakers placed on the floor near them. As they listened, they were asked to tap the floor with their right arm. The timing of the sound and the position of the speaker were varied, with the speakers getting further and further away. The illusion occurred when the sound was produced at twice the distance to the location where the subjects tapped. When the subjects' arms were then stimulated after they had tapped the floor, they tended to say that the distance of the two stimulation points was farther apart on their right arm than on their left arm. This creates a spatial inconsistency between the somatic senses and hearing. When experiencing these conditions, our perceptual system may be interpreting them as, *the sound is coming from where my right hand is tapping*, and *my right hand is tapping far away*. Further, if that is the case, we might also feel that *my right arm must be pretty long*.

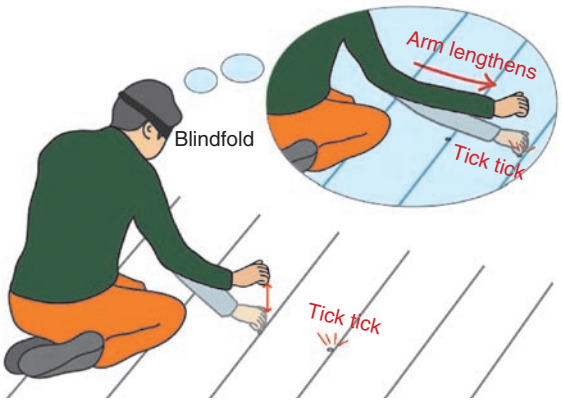


Fig. 3. Action sounds heard from a distance induce the illusion of a longer arm.

Experiments conducted to investigate this showed that, in fact, subjects perceived their right arm to be longer than it actually was. After experiencing this condition, objects presented on the right arm were

felt to be longer than those felt before the experience. The fact that the same objects that touched the skin of the arm were perceived to be longer after the experience suggests that the arm is recognized as being longer in the brain. The illusion occurred when the sound was produced at twice the distance to the location of the tapping, but not at four times the distance. It was also important that the participants actively tapped the floor themselves. Additionally, it was shown that the illusion occurs at an unconscious level. These findings show that the spatial position of sounds produced by our own bodily motions provide unconscious clues in how we perceive our own bodies.

Our recognition of our own bodies is constantly and flexibly being updated. This suggests that any bodily sensation can be produced by controlling it appropriately. Further, it should be possible to make the experience of virtual reality spaces more realistic by inducing such bodily sensations. These changes in body recognition are also thought to be related to how users perceive a tool as becoming part of their body when they become accustomed to using it. The extent to which bodily perception can be extended, spatially and temporally, is a topic we are considering for future research.

5. Future developments

We believe that bodily sensations are deeply related to the reality of experiences. Most experiences that involve no bodily sensations feel unsatisfying in some way. Most seeing or hearing sensations relate to phenomena that are distant from the body, but touch is said to be able to confirm the existence of what is being touched [9]. Further, what we recognize to be our own body is what we recognize as ourselves, which is the basis for having a concept of the self. A concept of the self is also necessary for understanding others, so bodily sensation is fundamentally important for our communication.

The fact that the sense of hearing contributes to perception of the body is extremely important for information and communications technology, which is a medium for communication. It is difficult to transmit bodily sensations such as touch, both technically and in terms of cost, but it is relatively easy to transmit sound. It seems that by arranging how sound is presented, it should be possible to convey experiences with a level of reality that accompanies bodily sensations. Elucidating the mechanisms of body awareness should contribute to realizing better and deeper means of communication in the future.

References

[1] M. Botvinick and J. Cohen, "Rubber hands "feel" touch that eyes see," Nature, Vol. 391, No. 6669, p. 756, 1998.

[2] K. C. Armel and V. S. Ramachandran, "Projecting sensations to external objects: evidence from skin conductance response," Proc. of R. Soc. B Biol. Sci., Vol. 270, No. 1523, pp. 1499–1506, 2003.

[3] N. Kitagawa and Y. Igarashi, "Tickle sensation induced by hearing a sound," Jpn. J. Psychon. Sci., Vol. 24, No. 1, pp. 121–122, 2005.

[4] N. Kitagawa, M. Zampini, and C. Spence, "Audiotactile interactions in near and far space," Exp. Brain Res., Vol. 166, No. 3-4, pp. 528–537, 2005.

[5] A. Tajadura-Jiménez, N. Kitagawa, A. Väljamäe, M. Zampini, M. Murray, and C. Spence, "Auditory-somatosensory multisensory interactions are spatially modulated by stimulated body surface and acoustic spectra," Neuropsychologia, Vol. 47, No. 1, pp. 195–203, 2009.

[6] N. Kitagawa and C. Spence, "Audiotactile multisensory interactions in human information processing," Jpn. Psychol. Res., Vol. 48, No. 3, pp. 158–173, 2006.

[7] V. Jousmäki and R. Hari, "Parchment-skin illusion: sound-biased touch," Curr. Biol., Vol. 8, No. 6, p. R190, 1998.

[8] A. Tajadura-Jiménez, A. Väljamäe, I. Toshima, T. Kimura, M. Tsakiris, and N. Kitagawa, "Action sounds recalibrate perceived tactile distance," Curr. Biol., Vol. 22, No. 13, pp. R516–R517, 2012.

[9] J. Watanabe, "Communication Research Focused on Tactile Quality and Reality," NTT Technical Review, Vol. 9, No. 11, 2011. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201111fa6.html

Norimichi Kitagawa

Senior Research Scientist, Human Information Processing Laboratory, NTT Communication Science Laboratories.

He received the Ph.D. degree from Tokyo Metropolitan University in 2003. He received a JSPS (Japan Society for the Promotion of Science) Research Fellowship for Young Scientists (Superlative Postdocs) and spent time at the University of Oxford, UK, in 2003. Upon his return, he took up a lectureship at Kanazawa Institute of Technology. He joined NTT Communication Science Laboratories in 2005. He is interested in how we perceive the world around us and how our perceptual systems process the information from different senses such as hearing, touch, and vision.

Efficient Mining Algorithms for Large-scale Graphs

Yasunari Kishimoto, Hiroaki Shiokawa, Yasuhiro Fujiwara, and Makoto Onizuka

Abstract

This article describes efficient graph mining algorithms designed for analyzing large-scale graph data such as social graphs. Graph mining is a technique to analyze the structure of graphs consisting of nodes and edges. We have developed efficient algorithms for two mining tasks: clustering and computing personalized PageRank, for large-scale graphs.

Keywords: graph mining, clustering, personalized PageRank

1. Introduction

One of the methods used to analyze big data is to handle it as graph data. Graph data consist of nodes and edges, where each edge connects two nodes (**Fig. 1**). Graph mining is a technique for discovering hidden relationships between various data by analyzing graph data. The amount of research on graph data has been increasing rapidly in recent years, and more services that generate graph data, for example, social networking services (SNSs) are being offered. Consequently, graph mining is becoming a major trend in big data analysis. However, efficient techniques for analyzing web-scale graph data have not been well established yet.

We have been conducting research to design

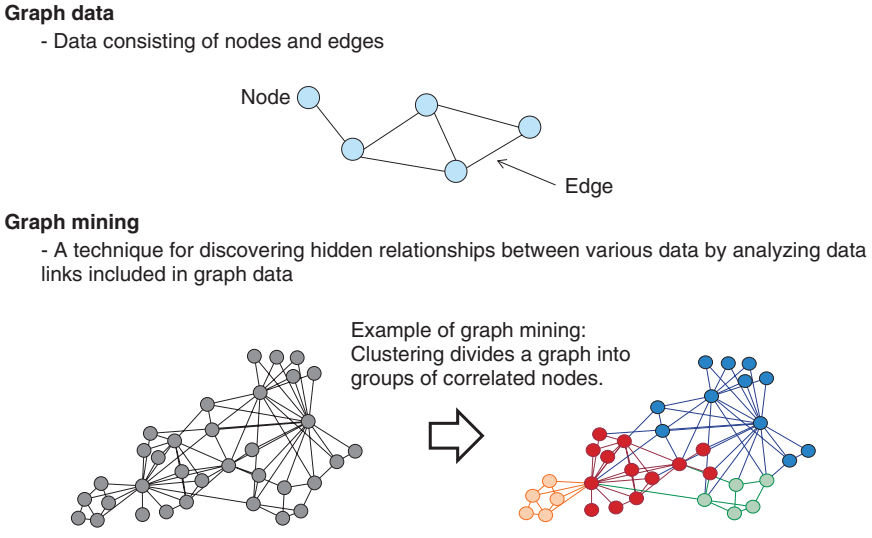


Fig. 1. Graph data and graph mining.

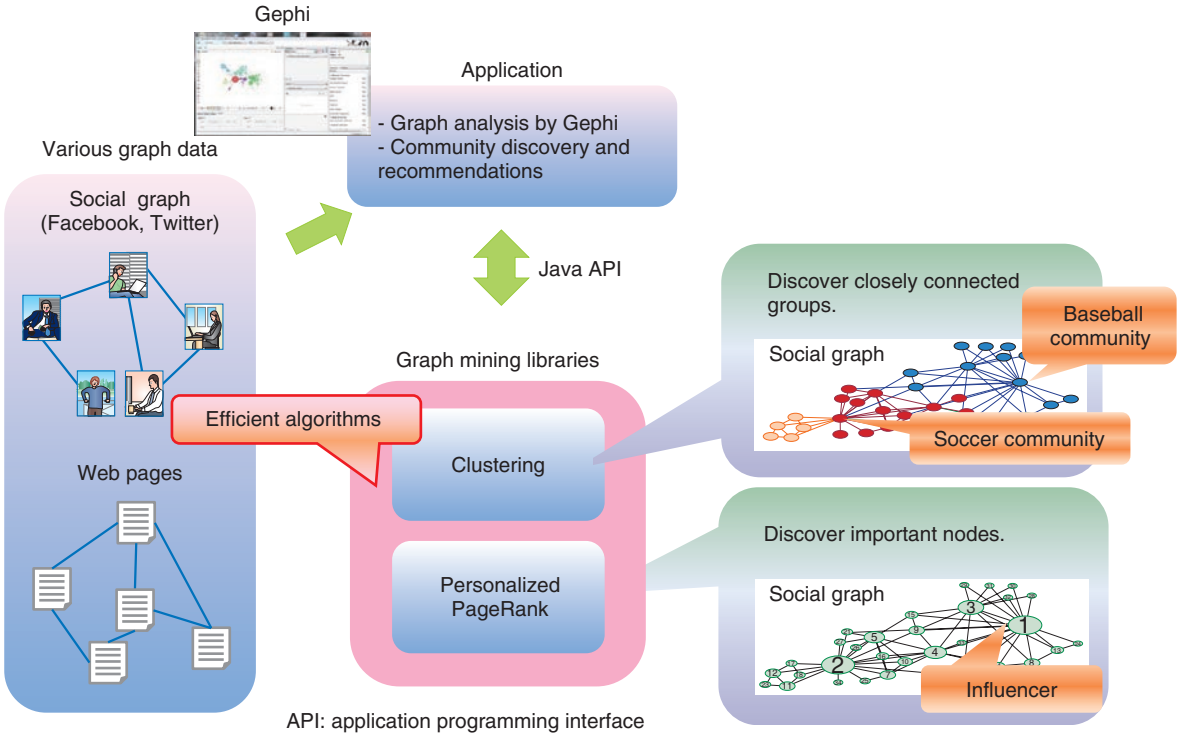


Fig. 2. Summary of graph mining techniques.

efficient algorithms for graph mining. In February 2013, we announced that we had developed the world's fastest algorithms for performing two techniques [1]. One is clustering, which involves grouping graph data by taking the density of edges between nodes into account. The other is computing personalized PageRank, which involves searching graph data for nodes with high importance values.

These new algorithms make it possible to analyze graph data significantly faster than conventional algorithms. For example, in clustering, conventional algorithms take 4–6 hours to analyze the SNS friendship relationship of 100 million persons. The response time is reduced to only 3 minutes by applying our new algorithm. This substantial reduction results in qualitative changes in analysis, and it provides many applications that bring opportunities for graph mining analysis. We also confirmed that in addition to its high efficiency, our method for clustering (see section 2) achieves a level of accuracy as high as that of the most accurate conventional method (for example, the Louvain method [2]). Conventional algorithms have been devised to speed up the response time at the expense of analysis accuracy. In contrast, our methods have achieved speed-up without loss of accuracy.

We have developed both Java libraries and Gephi plug-ins for the algorithms in order to make them widely available. Gephi is a graph analysis and visualization tool that is freely available; it is implemented in Java (**Fig. 2**).

2. Clustering algorithm

The efficiency of our clustering algorithm [3] is achieved by applying two approaches, as illustrated in **Fig. 3**.

The first approach is to optimize the computation order of nodes during clustering by using the statistics of the graph structure. The degree of a node is defined as the number of edges coming from the node. The computation starts from the node with the smallest degree. This design is based on the bottom-up clustering procedure in which two nodes connected by edges are merged step-by-step until the cluster quality of the graph no longer increases. The merge cost depends on the degree of the nodes; the higher the node degree, the more edges have to be referenced during merging. Consequently, the cost increases.

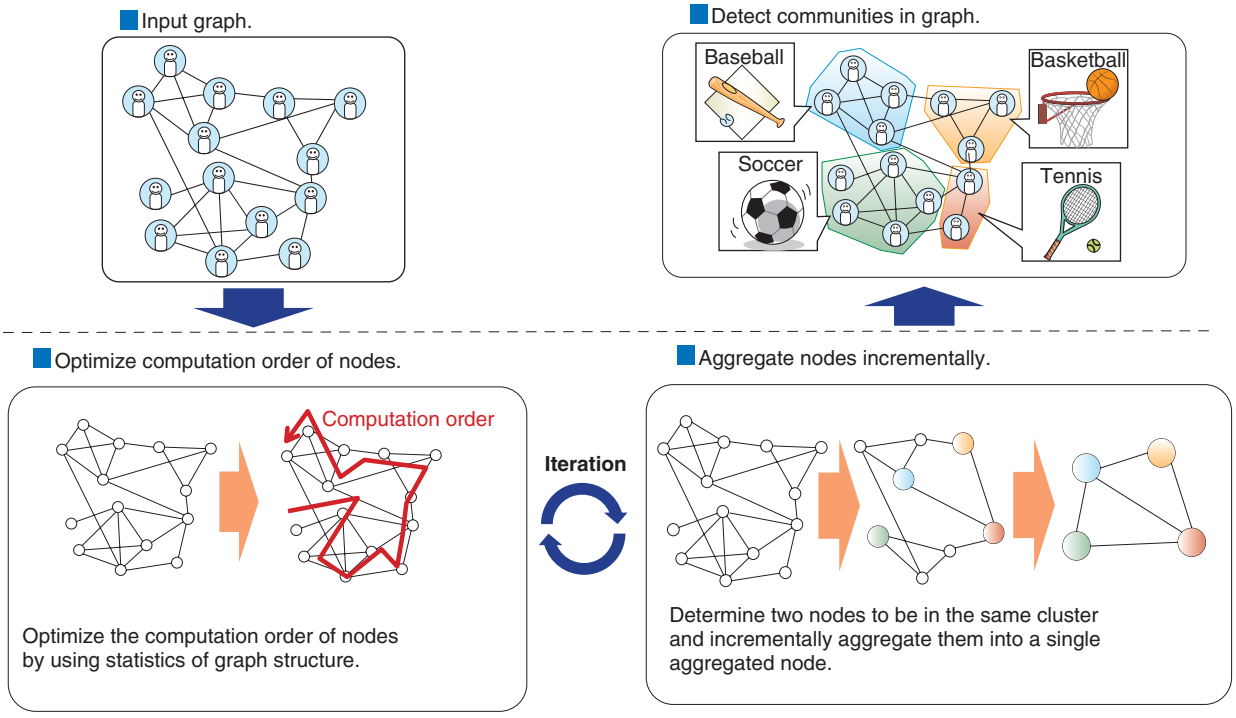


Fig. 3. Overview of efficient clustering algorithm.

The second approach is to incrementally aggregate two nodes that have been determined to be in the same cluster (or group) into a single aggregated node. This node aggregation is repeated until the cluster quality no longer increases. Since the number of nodes and edges is increasingly reduced through the repeated aggregation, it dramatically improves the response time of clustering. In addition, we introduce an incremental pruning technique for more efficient clustering. That is, if there is a node with only one edge, this node can be aggregated to the cluster of the single neighbor node without computing the increase in the cluster quality. General graph data have many such nodes, so this incremental pruning technique is very effective.

The use of these approaches means that our clustering algorithm performs from 10 to 60 times faster than previous algorithms.

3. Personalized PageRank algorithm

We developed an efficient personalized PageRank [4] algorithm for top-k search (**Fig. 4**) by applying two approaches.

The first approach is to efficiently compute the importance values, that is, the personalized PageRank (PPR) scores of nodes, by permuting rows and columns of the adjacent matrix of input graph data so as to increase the number of zero elements in the matrices, which are obtained by decomposing the adjacent matrix. Since any elements multiplied by zero become zero in the decomposed matrices, we can reduce the computation cost by increasing the number of zero elements.

The second approach is that instead of computing the exact PPR scores of all nodes, we efficiently identify top-k nodes with the highest PPR scores by estimating the upper bound of the PPR scores of nodes. The process of identifying top-k nodes is as follows. First, we estimate the upper bound of PPR scores of top-k candidate nodes. Next, we compute the precise PPR scores of the top-k candidate nodes. Then, we continue estimating the upper bound of PPR scores of other promising nodes. We can stop the estimation when the upper bound PPR scores of the promising nodes are lower than the PPR score of the current top-kth node. This means that it is not necessary to compute the precise PPR scores of all nodes in general, which reduces the total cost of PPR computation.

As a result, our personalized PageRank algorithm

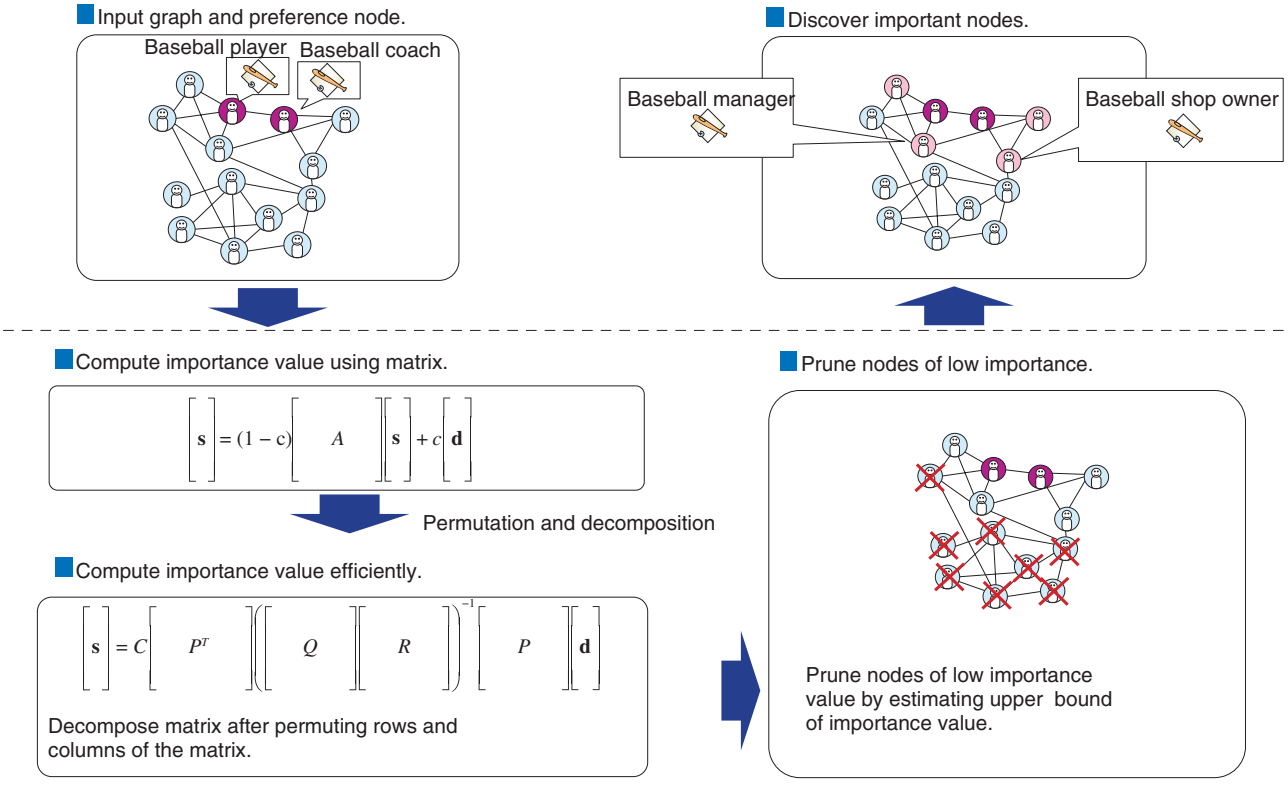


Fig. 4. Overview of efficient personalized PageRank algorithm.

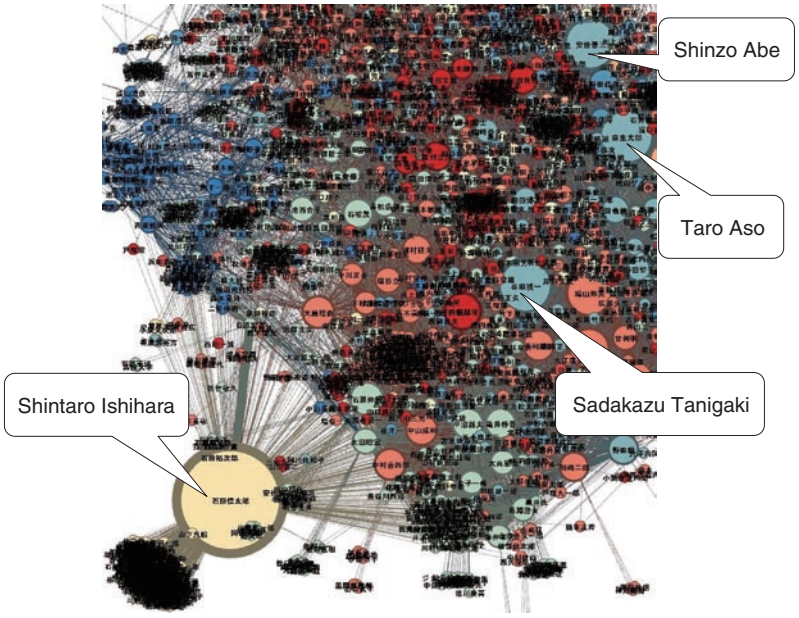


Fig. 5. Example of mining a social graph of politicians.

for top-k search is more than 50 times faster than conventional algorithms.

4. Example of graph mining application

A practical application of our graph mining algorithms is illustrated in **Fig. 5**. We obtained a social graph of the members of Japan's House of Representatives and their friends using Wikipedia. Then we applied our algorithms to the graph and analyzed the communities (clusters) of the members and their importance values (PPR scores). The politicians who belong to the same community have the same color in the figure, and those whose importance value is higher have a larger node. There were about 3,000 nodes in total.

We can see in the figure that the size of the *Shintaro Ishihara* node is the largest. This suggests that he is a very influential person. Meanwhile, the nodes for *Shinzo Abe, Taro Aso,* and *Sadakazu Tanigaki* have the same color. This suggests that they belong to the same community. In fact, they all belong to the Liberal Democratic Party.

5. Summary

We have developed the world's fastest algorithms for the graph mining tasks of clustering and computing personalized PageRank. We have also implemented both Java libraries and Gephi plug-ins for our algorithms to make the algorithms widely available. We implemented some approaches to make our algorithms highly efficient and successfully applied our algorithms to a social graph extracted from Wikipedia.

References

[1] NTT news release (in Japanese). http://www.ntt.co.jp/news2013/1302/130213b.html

[2] V. D. Blondel, J-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of Statistical Mechanics: Theory and Experiment, 2008:P10008, October 2008.

[3] H. Shiokawa, Y. Fujiwara, and M. Onizuka, "Fast Algorithm for Modularity-based Graph Clustering," Proc. of the 27th AAAI Conference on Artificial Intelligence (AAAI 2013), Bellevue, WA, USA.

[4] Y. Fujiwara, M. Nakatsuji, T. Yamamuro, H. Shiokawa, and M. Onizuka, "Efficient Personalized PageRank with Accuracy Assurance," Proc. of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2012), Beijing, China.

Yasunari Kishimoto
Research Engineer, Distributed Computing Technology Project, NTT Software Innovation Center.
He received the B.E. and M.E. degrees from Kyushu University, Fukuoka, in 1989 and 1991, respectively. He joined NTT in 1991 and studied directory systems, billing systems, and data mining. He is a member of the Information Processing Society of Japan (IPSJ).

Hiroaki Shiokawa
Researcher, Distributed Computing Technology Project, NTT Software Innovation Center.
He received the B.E. and M.E. degrees in computer science from the University of Tsukuba, Ibaraki, in 2009 and 2011, respectively. He joined NTT in 2011 and has been studying graph data management, graph mining algorithms, distributed computing, and databases.

Yasuhiro Fujiwara
Research Engineer, NTT Software Innovation Center.
He received the B.E. and M.E. degrees from Waseda University, Tokyo, and the Ph.D. degree from the University of Tokyo in 2001, 2003, and 2012, respectively. He joined NTT in 2003. His research interests include data mining, databases, natural language processing, and artificial intelligence. He is a member of IPSJ, the Institute of Electronics, Information and Communication Engineers, and the Database Society of Japan.

Makoto Onizuka
Distinguished Technical Member, NTT Software Innovation Center and Visiting Professor at the University of Electro-Communications.
He received the Ph.D. degree in computer science from Tokyo Institute of Technology in 2007. During 2000–2001, he was at the University of Washington, Seattle, WA, USA, where he worked on XML stream engines and database systems. His research focuses on cloud-scale data management and analytical processing.

Trends Concerning Standardization of OpenADR

Ryutaro Toji

Abstract

Automated demand response (ADR) technology is drawing worldwide attention alongside renewable energy technologies as a countermeasure against global warming and rising energy costs. In Japan, standardization of ADR has progressed rapidly as a promising power-saving measure since the Great East Japan Earthquake of March 2011, and Demand Response Interface Specification Version 1.0 was adopted by the JSCA (Japan Smart Community Alliance) Smart House/Building Standardization and Business Promotion Study Group organized by METI (Ministry of Economy, Trade and Industryof Japan) in May 2013. This article describes OpenADR 2.0, the international standard that forms the basis of the above-mentioned Japanese specification.

Keywords: automated demand response, smart community, OpenADR

1. Introduction

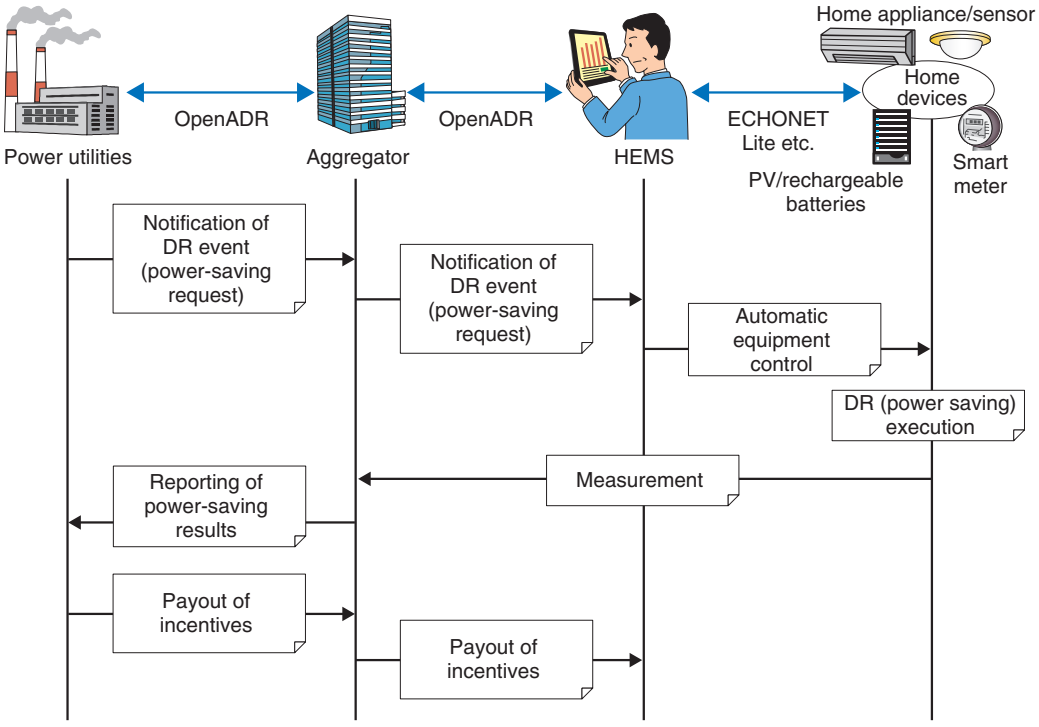
Demand response (DR) is defined as the practice of reducing peak power consumption based on utilities' demands for power saving and electricity users' (consumers') responses to them. For example, utilities or aggregators[*1] notify consumers of power saving incentives or temporary increases in electricity prices in accordance with the peak power demand. Then consumers restrain their power usage during peak times or shift it to off-peak times (**Fig. 1**). This helps to reduce power consumption during peak times.

Because the cost of storing electricity is very high, it is not practical to store electricity for peak time use, and utilities must therefore keep preliminary power generators to prepare for peak time power usage. Thus, from the standpoint of utilities, reducing peak power consumption by using DR has the same effect as cutting back on power-generation facilities and reducing fuel costs. In the United States, DR has been subjected to large-scale field tests around the country since the mid-2000s as a result of power shortages that have occurred because of restrictions on the construction of new power plants imposed by environmental regulations. These tests have confirmed that DR can reduce electricity demand by roughly 10–20%. In particular, automated demand response (ADR), whereby messages are exchanged between systems electronically and devices are controlled automatically using energy-management systems (EMSs), has been recognized as being more effective than manual DR, whereby notifications of DR events are sent to users via email, and consumers then control their devices manually. OpenADR 1.0 was developed as a message-exchange protocol for ADR by the Lawrence Berkeley National Laboratory in the USA. After some field tests and commercialization activities by several California power utilities, it was made public in 2009 (see **Fig. 2**).

2. Steps toward standardization of OpenADR

OpenADR 1.0 was originally just a local standard in the USA. However, it took a step closer to being an international standard after it was recommended by NIST (the National Institute of Standards and Technology) in the USA as one of the standards that should be complied with to achieve interoperability of smart grids. In 2009, NIST established the Smart Grid Interoperability Panel (SGIP) as a body

*1 Aggregator: a business operator acting as an intermediary between power utilities and groups of consumers that aggregates the amounts of megawatt power, namely the reducible power demand saved by consumers in exchange for incentives.



ECHONET: Energy Conservation and Homecare Network
HEMS: home energy management system
PV: photovoltaic

Fig. 1. Example of an ADR sequence.

promoting the standardization of smart grids, and at the behest of SGIP, the Organization for the Advancement of Structured Information Standards (OASIS[*2]) started developing standards for ADR. Referring to previously investigated documents such as the OpenADR 1.0 System Requirement Specifications by the Utility Communication Architecture International Users Group (UCAIug) and the results of investigations by the North American Energy Standards Board (NAESB), OASIS drew up two specifications—EI 1.0 (Energy Interoperation 1.0) and EMIX 1.0 (Energy Market Information Exchange 1.0)—as new international standards. Although these specifications can be applied to ADR as well as a wide range of business transactions in the electric industry, they are not necessarily sufficient from the viewpoint of implementation. Under those circumstances, the OpenADR Alliance was established in 2010 as an international standardization body for developing and promoting a practical ADR standard (namely, OpenADR 2.0) and providing certification to approved products. The OpenADR Alliance established an interoperable standard by extracting specifications required by ADR from EI 1.0 and EMIX 1.0 and supplemented insufficient points regarding implementation. A set of specifications of this kind is called a Profile. OpenADR 2.0 Profile A (2.0a), which targets control of relatively simple devices such as intelligent thermostats, was made public in August 2012, and OpenADR 2.0 Profile B (2.0b), which targets full-blown ADR services offered by aggregators, was made public in July 2013. Profile 2.0b incorporates 2.0a and offers higher functionality than its predecessor. Therefore, this article focuses on 2.0b.

3. Overview of OpenADR 2.0

OpenADR stipulates data models for message exchange and communication protocols needed when

*2 OASIS: an international non-profit organization promoting standardization of XML (extensible markup language) based electronic transactions between businesses.

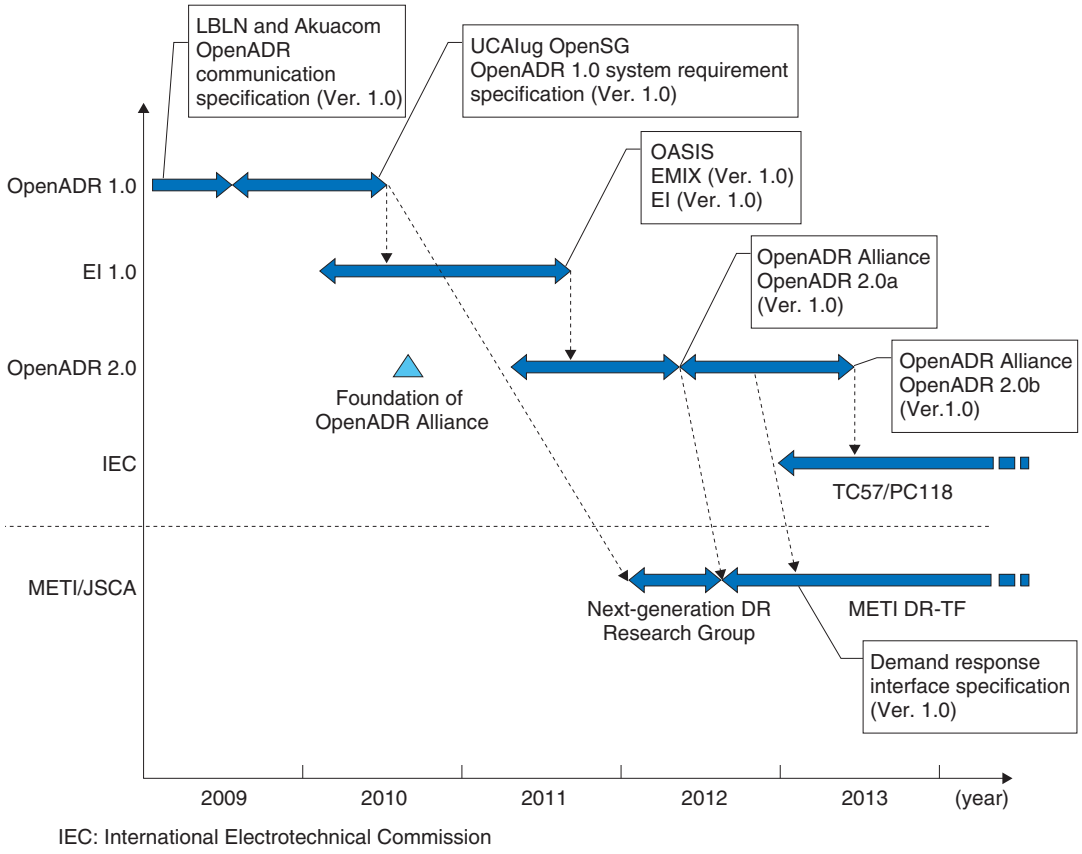
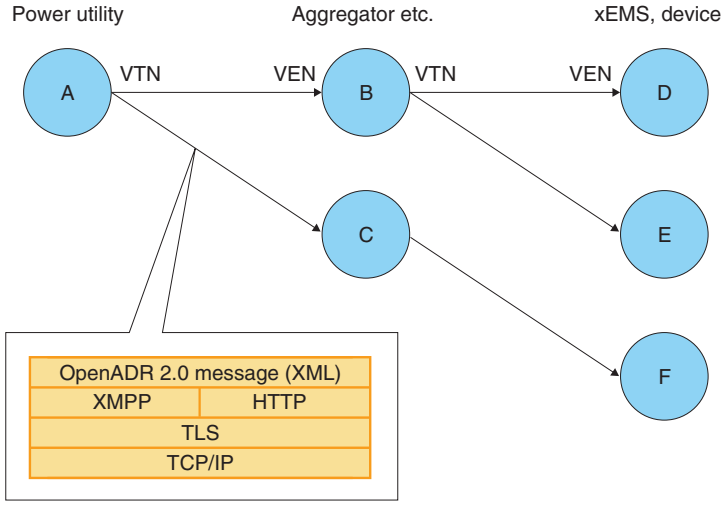


Fig. 2. Steps involved in standardization of OpenADR.

information is exchanged between *actors*, namely business entities such as power utilities and consumers, involved in ADR services. As shown in **Fig. 3**, the actors are modeled in the form of virtual top nodes (VTNs), which send messages, and virtual end nodes (VENs), which receive messages. Power utilities and consumers correspond to VTNs and VENs, respectively. Aggregators are actors that double as VTNs for consumers and VENs for utilities; they can be set up in multiple stages. A VTN is generally implemented as a server system called a demand response automation server (DRAS), while a VEN usually corresponds to "x energy management systems" (xEMSs), for example, a BEMS (business EMS) or HEMS (home EMS), and devices such as smart meters.

VTNs and VENs are assumed to exchange messages via the Internet (IP (Internet protocol) network). The transport mechanism is defined as either of two types of communication models, namely *Pull* and *Push*. In a Pull model, a VEN polls a VTN periodically and receives messages from it. By contrast, messages are sent from a VTN to a VEN in the Push model. The protocols that support these models are standard HTTP for the former, and XMPP (extensible messaging and presence protocol), which enables bidirectional communication used for instant messaging, for the latter. Messages are defined in XML, and their data structure is defined in XSD (XML schema definition). In addition, a standard level of security using TLS (transport layer security) and a high level of security combining an XML signature are prescribed for securing messages.

The services provided by OpenADR 2.0b are listed in **Table 1.** Four services are prescribed: EiRegisterParty, which reciprocally registers the VTNs and VENs that become communication parties; EiEvent, which transmits details of DR events; EiReport, which sends measured and recorded data related to DR events, for example, power consumption; and EiOpt, which manages acceptances and refusals concerning DR events. Meanwhile, in OpenADR 2.0b, an extremely wide range of required and optional



HTTP: hypertext transfer protocol
TCP/IP: transmission control protocol/Internet protocol

Fig. 3. Communication model for OpenADR 2.0.

Table 1. Services provided by OpenADR2.0.

Service	Summary
EiRegisterParty (Registration)	VTN and VEN are mutually registered; information required for reciprocating messages is mutually exchanged.
EiEvent (DR events)	VTN notifies VEN of DR events, updates notification contents, and cancels notifications. Valid period of events and event contents are shown. Various event contents (such as price information, assignment of load reductions, load control, and rechargeable-battery control) are defined.
EiReport (Reporting/feedback)	Instantaneous and accumulated values of measurement results (such as power consumption and voltage at VEN) are reported. Prior to the reporting, information concerning respective reporting capabilities is mutually exchanged.
EiOpt (Acceptance/modification)	Schedules of acceptances and rejections regarding DR events are transmitted from VEN to VTN.
(Remarks)	Although services such as registration of resource equipment controlled by ADR (EiEnroll), management of available resources (EiAvail), and cooperation with markets (EiMarketContext) were initially going to be defined in OpenADR2.0, they were omitted from Profile B since the present specifications were judged sufficient for actual usage.

parameters are defined so that a variety of ADR services can be created. When installing OpenADR 2.0b, it is necessary to set each parameter in line with the contents and operation of services.

Although it was initially planned that OpenADR 2.0 would evolve into Profile C, hearings conducted with various power utilities indicated that the current specifications are sufficiently practical, so Profile C was halted. It is possible, however, that services not covered by Profile B will be added in the future, for example, management of devices that provide DR resources and coordination with markets.

4. Status of standardization concerning ADR in Japan

In contrast to the worldwide boom in smart grids that was triggered by the energy policies promoted by President Obama in the USA, some skepticism towards this worldwide trend was initially exhibited in Japan, where the energy supply was considered to be stable. However, following the Great East Japan Earthquake of March 2011, DR started to draw attention as a promising countermeasure to power shortages. Consequently, proving tests and test services—

starting with field tests in four regions of Japan performed by the Japan Smart Community Alliance (JSCA)—were started in various regions. These tests were, however, limited to manual DR or cooperation with systems through independent communication protocols. To apply ADR as a genuine social infrastructure, the standardization of an interface between business operators (utilities and aggregators) and consumers was urgently required.

Given those issues, METI set up the Next-generation Demand Response Technical Standardization Research Group under the JSCA Smart House/Building Standardization and Business Promotion Study Group in June 2012, and surveys on use cases of DR in Japan and technical investigations on OpenADR were carried out. In their final report in September 2012, this research group confirmed the validity of adopting OpenADR 2.0 as the basis of ADR standards in Japan.

On the basis of this report, METI then set up the Demand Response Task Force (DR-TF) under the above-mentioned study group, and the DR-TF drew up the technical specifications required for DR communications between energy suppliers (power utilities) and energy users (consumers and aggregators) in Japan. This specification was adopted in May 2013 and is called Demand Response Interface Specification Version 1.0.

This specification became somewhat transitional in various ways in order to ensure that it was ready for application in the summer of 2013. For example, the specification is applied only to the interface between the power utilities and the aggregators. The interface between aggregators and xEMSs was deferred, even though it would be more effective to standardize it. Moreover, the specification should be called a *minimum subset* of 2.0b, since it consists of 2.0a and a very small part of version 2.0b because of the considerable delay in formulating specification OpenADR 2.0b itself. However, ADR services themselves are still in the earliest stages of development throughout the world, and there is a lot of room to improve them by refining them through actual use. METI is at the point of thoroughly reviewing Version 1 of the specification in light of the results of verification tests and service applications implemented in 2013.

5. OpenADR at NTT R&D (Research and Development)

At NTT Network Technology Laboratories, studies on the *Smart Community* have been ongoing since the autumn of 2011, and R&D on a Smart Community Platform (SCPF) is now underway. From the early stage of development, ADR has been viewed as a fundamental technology, and OpenADR 1.0 was adopted for SCPF Ver. 1.0, which was developed in October 2012. It appears more than likely that SCPF Ver. 1.0 was the initial implementation of OpenADR 1.0 in Japan. SCPF Ver. 1.0 was applied to a field test on ADR carried out from March to August 2013 at a company house of NTT EAST. In parallel with this development, as well as joining the OpenADR Alliance in September 2012, NTT Network Technology Laboratories joined the above-mentioned research group and DR-TF organized by METI, and has been contributing knowhow based on the experience gained in implementing the SCPF. Smart Community Platform Ver. 1.6 (SCPF Ver. 1.6) was developed in June 2013 as a B2B2C (business-to-business-to-consumer) platform for power utilities and aggregators; it can provide ADR services via the cloud based on OpenADR 2.0 (**Fig. 4**). SCPF Ver. 1.6 was certified by the OpenADR Alliance to conform to OpenADR 2.0a in July 2013 and to conform to OpenADR 2.0b in October 2013. It is the first OpenADR 2.0 certified product in Japan. In the meantime, SCPF Ver. 1.6 is being implemented in accordance with Demand Response Interface Specification Version 1.0, which was recommended by METI as mentioned previously, and it has been adopted as a core system for ADR verification tests at Waseda University. This test is subsidized by METI and is being conducted by Waseda University in order to verify the feasibility and interoperability of the Demand Response Interface Specification by applying it to ADR communications between actual power utility and aggregators' ADR systems.

6. The future of OpenADR

As described in the above sections, OpenADR 2.0 is being adopted in Japan and other countries as an international standard for implementing ADR and is being verified through proving tests in those countries. At the present time, OpenADR 2.0 is being positioned as the de facto or forum standard. However, it has been decided by committees TC57 and PC118 of the IEC (International Electrotechnical Commission) that work will commence on harmonizing the IEC's common information model IEC61850, namely, a power-system information model, with OpenADR 2.0b. It can thus be said that OpenADR is on its way to becoming the de jure standard. This

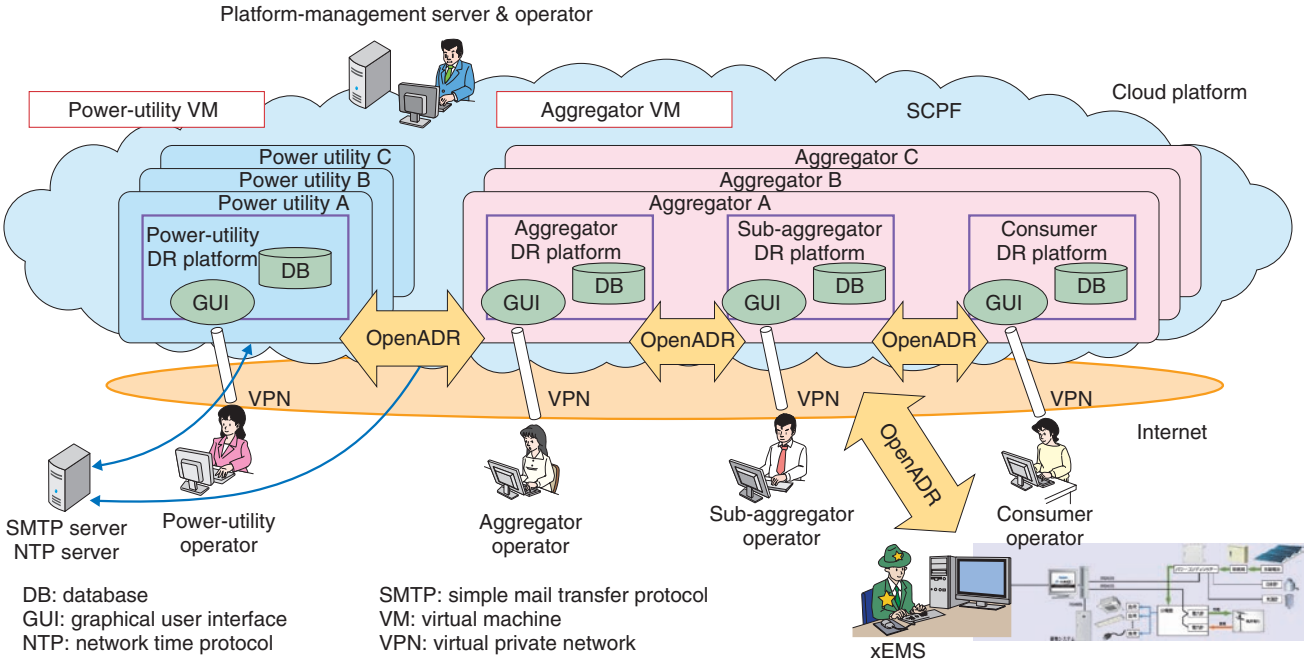


Fig. 4. Overview of Smart Community Platform (Ver. 1.6).

standardization work is predicted to take several years, so in the meantime, OpenADR 2.0b is being positioned as a PAS (publicly available specification).

Moreover, if we look into the future, ADR is not simply a way of providing power-saving services. Essentially, it is about reinventing the electricity infrastructure by incorporating power-saving activities and renewable energy on the consumer side into conventional power distribution, which has been a one-way street from power utility to consumer. From that viewpoint, it is sufficient to say that OpenADR 2.0b will be upgraded and expanded from now on as that reinvention continues.

References

[1] "2008 Assessment of Demand Response & Advanced Metering Staff Report," Federal Energy Regulatory Commission.
[2] "Demand Response Program for Residential Customers in the United States—Evaluation of the Pilot Programs and the Issues in Practice–," CRIEPI REPORT Y10005, Central Research Institute of the Electric Power Industry, 2011 (in Japanese).
[3] OpenADR 2.0 Profile Specification B Profile (Rev. 1.0), OpenADR Alliance, 2013.
[4] NTT news release (in Japanese). http://www.ntt.co.jp/news2013/1307/130722a.html

Ryutaro Toji
Senior Research Engineer, Supervisor, Network Technology Project, NTT Network Technology Laboratories.
He received the M.S. degree in coordinated science from the University of Tokyo in 1987. After joining NTT Communications and Information Processing Laboratories in 1987, he engaged in R&D of dialogue systems using speech recognition technology, computer telephony integration systems for call centers, and a network-based management platform for multi-application smart cards, which was deployed for the Basic Resident Registration Card in Japan. He was a General Manager in the Platform Services Department and the Innovative Financial Systems Department of NTT Communications from 2005 to 2010. Since 2011 he has been engaged in R&D of the Smart Community Platform using Automated Demand Response technology at NTT Network Technology Laboratories.

Enhancing the Reliability of Aerial Iron Fittings (Span Clamps and Outdoor Wire Anchors)

Abstract

In this article, we introduce how to enhance the reliability of aerial iron fittings (span clamps and outdoor wire anchors). This is the twentieth in a bimonthly series on the theme of practical field information on telecommunication technologies. This month's contribution is from the Materials Engineering Group, Technical Assistance and Support Center, Maintenance and Service Operations Department, Network Business Headquarters, NTT EAST.

Keywords: countermeasure, corrosion, telecommunication facilities

1. Introduction

Despite the coming of the optical era, metal, concrete, plastic, and other materials are still being used in telecommunication facilities. In the Materials Engineering Group, we are exploring the causes of failures unique to these facilities that have been brought on by the deterioration of these materials and are working to prevent the reoccurrence of such failures.

In this article, we present a case study of corrosion promotion peculiar to coastal areas in span clamps and outdoor wire anchors (aerial iron fittings) used in telecommunication facilities, and we introduce products to counter such corrosion-based deterioration.

2. Problems in span clamps and outdoor wire anchors leading to hanging drop lines

A span clamp and outdoor wire anchor are needed to drop an NTT telecommunication cable into a customer's home. If a span clamp should come apart from a messenger wire or a self-supporting type of telecommunication cable, or if an outdoor wire anchor should come apart from the span clamp, the drop line will hang loose, which will create a facility fault. A loose-hanging drop line can then come into contact with passing pedestrians or vehicles and cause an accident. It is therefore necessary to prevent such accidents by ensuring that these components do not separate from the equipment.

The separation of a span clamp or outdoor wire anchor can be caused by (1) loosening of the nut on the span clamp, (2) incorrect placement of the separation-prevention washer (**Fig. 1**), or (3) corrosion and subsequent deterioration of an iron fitting. Conditions (1) and (2) can be prevented by properly installing the components, but deterioration caused by corrosion (3) cannot be prevented by proper installation work alone.

3. Case study of corrosion and deterioration in span clamp and outdoor wire anchor

The main components of a span clamp are (1) support plates, (2) a hook bolt, (3) an anti-separation washer, and (4) a nut (**Fig. 2**). Support plates are made of zinc alloy, but components (2)–(4) are made of steel with zinc plating applied to their surfaces. Since zinc is more resistant to corrosion than steel (iron), the zinc-alloy support plates as well as the other components plated with zinc are thought to be sufficiently resistant to corrosion provided that they are not located in environments that are particularly conducive to corrosion such as areas with briny air or hot springs.

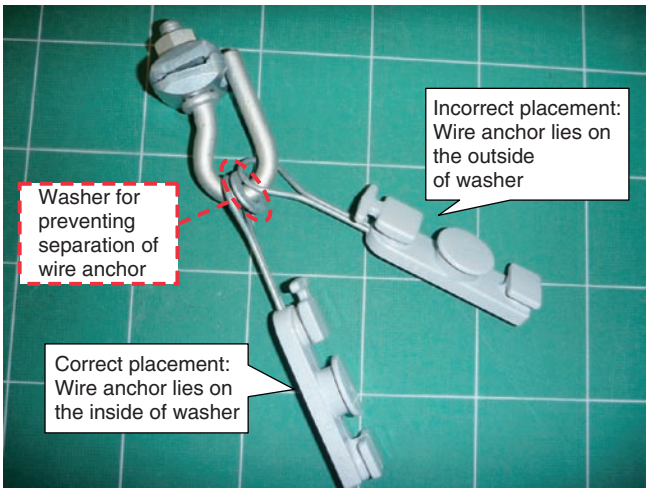


Fig. 1. Incorrect placement of washer for preventing separation of wire anchor.

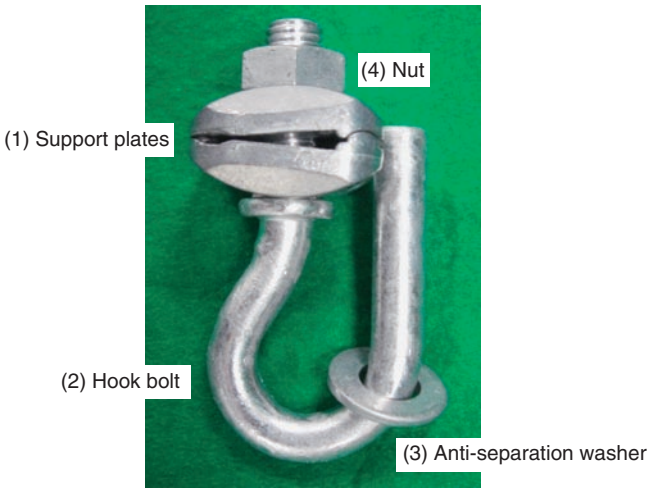


Fig. 2. Configuration of span clamp.

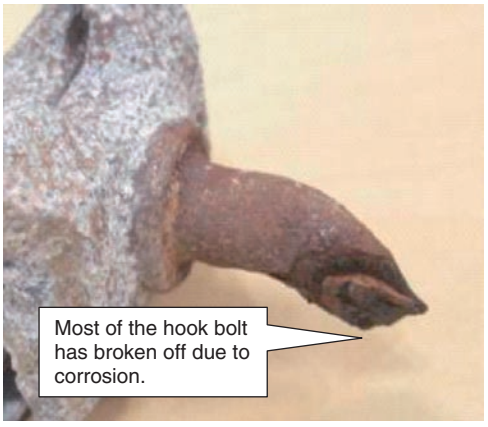


Fig. 3. Span clamp with broken hook bolt.

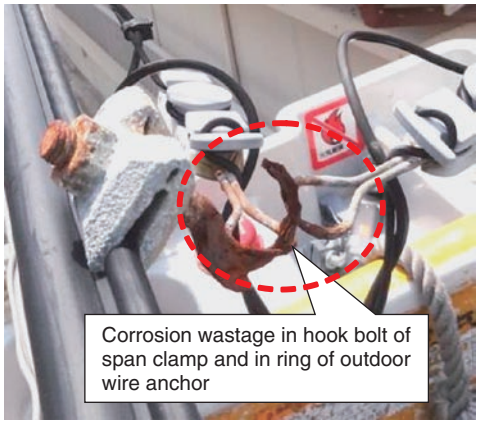


Fig. 4. Corroded span clamp and outdoor wire anchor.

A rusted and broken span clamp is shown in **Fig. 3**. It was left at a location in a coastal area where a drop line was hanging loose. This span clamp had been installed about 20 years earlier. It can be seen that the hook bolt is highly corroded and that most of it had broken off. The diameter of the hook bolt was approximately 9 mm, and its surface had originally been plated with a layer of zinc approximately 30 µm thick. Simulations done using commonly reported corrosion speeds for zinc plating and steel in coastal areas[*1] reveal that it would take about 50 years for corrosion to advance from the zinc-plating layer to a level deep enough to make the hook bolt break away. However, as stated above, this span clamp with the broken hook bolt had been installed only about 20 years earlier. Therefore, we continued our investigation in order to find the cause of this accelerated speed of corrosion in steel.

A corroded span clamp and outdoor wire anchor are shown in **Fig. 4**. These were installed in the same area as the above span clamp whose hook bolt broke off due to corrosion. Here, the corrosion has not yet reached the point at which the hook bolt would break

*1 Source for zinc plating: Japanese regional exposure test data (Corrosion Prevention Data Book, p. 92, Corrosion Prevention Handbook, p. 291—in Japanese).
Source for steel: Atmospheric Corrosion (Introduction to Material Environments, p. 160—in Japanese).

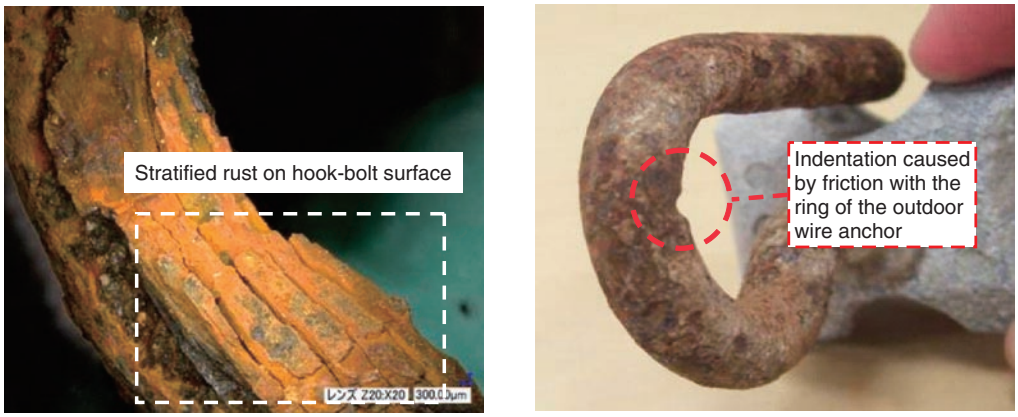


Fig. 5. Stratified rust on hook bolt (left) and friction indentation (right).

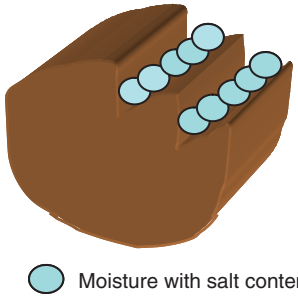


Fig. 6. Corrosion promotion scenario.

off, but it is clear that the span clamp and the ring fitting of the outdoor wire anchor are in a state of significant corrosion. Fragile, stratified rust covers the surface of the hook bolt, and a friction indentation can be seen at the point where the bolt comes into contact with the ring of the outdoor wire anchor (**Fig. 5**). A hook bolt that is not corroded has a cylindrical shape that causes rainwater or airborne sea salt to run off it, preventing it from collecting on the surface of the hook bolt.

However, if stratified rust should form on the surface of the hook bolt, water that contains salt—a corrosion-promotion factor—will collect and possibly lead to corrosion (**Fig. 6**). In addition, fragile, stratified rust can easily suffer from frictional wear due to friction between the bolt and the ring of the outdoor wire anchor. We therefore concluded that the generation of fragile, stratified rust and the frictional wear caused by friction with the corroded ring of the outdoor wire anchor were factors that most likely caused the hook bolt on the span clamp to break off about 20 years after installation.

4. Study of countermeasure products

As stated above, two factors are the main causes of the hook bolt breaking away from the span clamp: corrosion-induced deterioration of the material itself and frictional wear caused by friction between the hook bolt and the ring fitting of the outdoor wire anchor. With the aim of preventing the reoccurrence of such problems, we investigated ways of improving the corrosion resistance and abrasion resistance of span clamps and the ring fitting on outdoor wire anchors. We also took construction costs into account and set two conditions: (1) any enhanced products must be just as easy as existing ones to install, and (2) they must be manufacturable using nearly the same process as existing products.

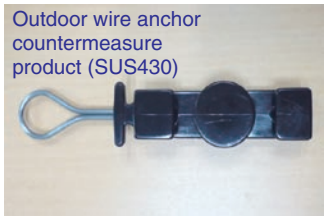
Specifically, we began by examining anticorrosion specifications that could be achieved by using a method for which manufacturing costs would not be greatly different from that of existing products. From that examination, we decided to investigate two types of countermeasures: anticorrosion techniques using a surface coating, for which anticorrosion could be achieved in existing products in a post-manufacturing process, and highly anticorrosive components using highly anticorrosive materials that could be manufactured on the same manufacturing line as existing products.

First, for the anticorrosion countermeasure using a

Table 1. Evaluation of prototype products.

Anticorrosion technique	Anticorrosion material	Corrosion resistance	Wear resistance	Ease of installation (shock resistance)	Workability
Surface coating	Zinc-aluminium alloy plating (Zn-5%Al)	×	×	◎	○
	Thermally sprayed film (zinc alloy: 100 μm)	△	×	◎	○
	Powder coating (PET)	◎	◎	×	◎
Highly anticorrosive components	Stainless steel SUS430	◎	◎	◎	○

Component	Conventional product	Countermeasure product
Hook bolt	Carbon steel for cold heading	SUS430 (stainless steel)
Nut		
Washer	Rolled steel for general structures	
Base	Zinc-alloy diecast	



Component	Conventional product	Countermeasure product
Ring fitting	Hard steel wire	SUS430 (stainless steel)
Base	Polyvinyl chloride resin	

Fig. 7. Comparison of component quality between conventional and countermeasure products.

surface coating, we prepared three prototype anticorrosion surface coatings by (1) using zinc-aluminium alloy plating, (2) thickening a surface zinc coating by thermal spraying, and (3) using a powder coating made of polyethylene terephthalate (PET) resin. Next, for the anticorrosion measure by using highly anticorrosive components, we prepared a prototype product using stainless steel (SUS430). We then evaluated the characteristics of each of the above prototype products in order to find the product with the best corrosion resistance, frictional-wear resistance, and ease of installation. The results are listed in **Table 1**.

We concluded from the evaluation that the prototype product using stainless steel components (SUS430) rather than the conventional material gave superior results.

A comparison of component quality between conventional products and the proposed countermeasure products is shown in **Fig. 7**. The washer of the span clamp and the base of the outdoor wire anchor of the countermeasure products are black to increase their visibility to maintenance personnel.

5. Conclusion

Corrosion of components used for telecommunication facilities can easily progress in certain regions such as coastal areas. We introduced a case study of deterioration in which the hook bolt of a corroded span clamp broke off due to friction with the ring fitting of an outdoor wire anchor. To prevent the reoccurrence of such deterioration, we improved the corrosion resistance and abrasion resistance of

conventional span clamps and outdoor wire anchors, thereby achieving products with long-term reliability.

Span clamps and outdoor wire anchors are components that are not recorded in design drawings, which make them difficult to maintain and inspect. At the same time, as facility maintenance personnel age and subsequently retire, their numbers in the years to come will most likely decrease, as it is expected to be difficult to replace all of them because of the declining population. The need is therefore strong for maintenance-free facilities. At the NTT EAST Technical Assistance and Support Center, we plan to conduct studies on efficiently maintaining and managing not only aerial iron fittings but underground iron fittings too.

Report on NTT Communication Science Laboratories Open House 2013

Akinori Fujino, Toshitaka Kimura, Kazushi Maruya, Marc Delcroix, and Hiroaki Sugiyama

Abstract

Open House 2013 was held in June at NTT Communication Science Laboratories in Keihanna Science City, Kyoto. Over 1000 people visited the facility on June 6 and 7 to enjoy 6 talks and 30 exhibits introducing our latest research activities and efforts in the fields of information and human sciences. This article reports on the main activities conducted during the open house.

Keywords: information science, human science, big data analysis

1. Introduction

At NTT Communication Science Laboratories (NTT CS Labs), we aim to build a new technical infrastructure connecting *people* and *information* and are therefore studying aspects of both human and information sciences to create innovative technologies and discover new principles. Branches of NTT CS Labs are located in Kansai Science City (Seikacho, Kyoto) and Atsugi City, Kanagawa. These laboratories deal with the most fundamental research targets in the fields of human and information sciences in the NTT laboratories.

The open house of NTT CS Labs has been held annually with the aim of introducing the results of the Labs' basic research and innovative leading-edge research to both NTT Group employees and visitors from corporations, universities, and research institutions who are engaged in research, development, business, and education.

This year, the event was held at the NTT Keihanna Building in Kansai in the afternoon of June 6 and all day on June 7, 2013. A total of 1140 visitors attended the event over the two days. Through talks and exhibits, we introduced our latest research results and explained how they were expected to affect future technologies and academic progress. In the special category of *big data analysis*, we presented exhibits organized by NTT CS Labs and other NTT laboratories. This article summarizes the event's research talks and exhibits.

2. Keynote speech

The open house started with a speech by the Director of NTT CS Labs, Dr. Eisaku Maeda, entitled "Cultivate trees that will bear fruit—Building a technical infrastructure that connects people and information—" (**Photo 1**).

Photo 1. Dr. Eisaku Maeda, Director of NTT CS Labs, giving his keynote speech.

Photo 2. Research presentation by Dr. Takaaki Hori.

Photo 3. Research presentation by Dr. Masaaki Nagata.

Since the start of the 21st century, both the volume and nature of the information that people have to deal with have changed greatly. Examples include ubiquitous information as well as that from sensor networks, cloud computing, and big data. The devices used for accessing communication networks have also largely shifted from desktop computers and cell phones to tablets and smartphones. Dr. Maeda talked about the research policy whereby, in the midst of these dramatic changes in the information environment, NTT CS Labs would study both mathematical principles and actual data to build a new technical infrastructure connecting *people* and *information*. He then introduced reverberation control, robust media search, question answering, and information science on material perception as successful research results produced by NTT CS Labs and applied in today's society. He mentioned that the initial research themes led us to new services and discoveries not even imagined when work started on the themes over a decade ago, and he emphasized that it was very important to cultivate research themes over such long periods of time.

3. Research talks

Four of the talks highlighted recent notable research results and high-profile research themes:

- "Towards a computer that can recognize everyone's conversations—Research on multi-speaker conversational speech recognition: past, present and future—", Dr. Takaaki Hori, Media Information Laboratory
- "The reality of the body—How we perceive our own body—", Dr. Norimichi Kitagawa, Human and Information Science Laboratory
- "Innovative user experiences brought by statistical machine translation—Breaking the language barriers of technical information", Dr. Masaaki Nagata, Innovative Communication Laboratory
- "Extracting hidden information from speech and audio signals—Generative modeling approach to speech and audio signal processing—", Dr. Hirokazu Kameoka, Media Information Laboratory

Each presentation introduced some of the latest research results and provided some background and an overview of the research. All of the talks were very well received.

Dr. Hori introduced the technology developed at NTT CS Labs for automatically recognizing conversational speech in meeting situations, and he discussed current problems related to conversational speech recognition and the future developments we can expect as more progress is made in this technology (**Photo 2**). Dr. Nagata's talk on innovative user experiences achieved through statistical machine translation focused on NTT CS Labs' latest research results on statistical machine translation and its automatic evaluation, in particular, between English and Japanese, which is one of the most difficult-to-translate language pairs in the world (**Photo 3**).

Photo 4. Research exhibits in *big data analysis* category.

4. Research exhibits

The open house featured 23 exhibits displaying NTT CS Labs' latest research results, and these were classified into the categories *innovative computing*, *media and communication*, and *human science*. The open house also included a special *big data analysis* category that consisted of two exhibits from NTT CS Labs and four from other NTT laboratories (**Photo 4**). The special category also included an exhibit summarizing the research directions that NTT laboratories were focusing on and the real-world problems being tackled in the fields of machine learning and data science.

Each exhibit was housed in a booth and was presented using slides on a large-screen monitor or hands-on demonstrations, with researchers explaining the latest results directly to visitors. The following list summarizes the research exhibits.

Big data analysis

- Challenges with big data—Recent big data analysis at NTT laboratories—
- Understanding stream data from essential elements—Feature selection for high-dimensional time series data—
- Smarter instant analysis of "Current" with big data—Streaming analysis with distributed online machine learning—
- Efficient graph mining techniques for big data—Fast algorithms for large-scale graphs—
- Network anomaly detection using big data analysis—Analyzing network failure/cyber attacks and their root causes—
- What is the author like? —Estimation of microblog user attributes using microblog structure—
- Massive trajectory data analysis and visualization—Mobility pattern analysis using heterogeneous data—

Innovative computing

- Efficient environmental monitoring—Correlated data gathering for sensor network—
- Exposing your whereabouts safely—Location privacy by using pseudonym exchange—
- Quantum information exchange employing optics—Optimal entanglement generation protocol with laser light—
- Making proper randomness—Physical random numbers generated by laser light—
- A visual language for advanced programming—Viscuit with new functions for advanced programming—
- Automatic grammatical analysis of English—Syntactic parsing based on statistical grammar induction—

Media and communication

- How can we overcome language barriers? —Statistical machine translation from foreign languages to Japanese—
- Yu bi Yomu: Reading experience based on trailing—A new method for reading dynamic texts using finger movements—
- Ways to support non-native speakers in a conference call—How transmission lags influence multilingual communication—
- Seeing into the mind in the twinkle of a smile—Analyzing expression/perception of interlocutors' empathy—
- Speech separation with collaborative recorders—Probabilistic fusion of different recording devices—
- Transcribing every word whoever speaks—Speech recognizers robust to casual speech variations—
- Transcribing known and unknown sounds—Bayesian semi-supervised audio event recognition—
- Instance search from large video collections—Video search technology with visual examples—
- Finding picture books suitable for a child—Graph-based similar picture book search with

weighted child words—

Human science

- Mystery of child word learning order—Cross-linguistic universality of child word learning periods—
- How children perceive phonemically similar words—Perceptual development of Japanese phonemes in children—
- When breath meets music—Playback system synchronizing phrases with respiration phases—
- Speak like a native—Speech rhythm control by non-negative temporal decomposition—
- How we feel fatigue and force—Motor perception influenced by delayed visual feedback—
- Future alters past—Postdictive processing in vision—
- Visual magic: changing an object just by watching—Visual illusion reveals neural codes for objects in the brain—
- Hearing the body—Action sounds recalibrate the perceived tactile distance—

"Yu bi Yomu: Reading experience based on trailing" introduced and demonstrated a novel method for reading digital texts on computers with a touch panel (**Photo 5**). With this method, the text is barely visible on the computer monitor. When a reader traces his/her finger across the monitor, the characters in the location the finger touches gradually appear and then disappear. This dynamic text presentation based on finger movement can provide a richer impression while reading.

"Hearing the body" introduced a phenomenon whereby a simple auditory trick was able to induce an illusion that made a participant feel that his/her arm was elongated. This exhibit included some demonstrations that showed how people were able to sense their own body through somatosensation, vision, and hearing (**Photo 6**).

Photo 5. Introduction of "Yu bi Yomu," a novel method for reading digital texts.

Photo 6. Visitor trying an illusion that makes him feel that his arm is elongated.

5. Invited Talk

This year's event also featured an invited talk by Prof. Masato Wakayama, Director, Institute of Mathematics for Industry, Kyushu University, which was entitled "New technology from/with mathematics and new mathematics from industry" (**Photo 7**).

He first talked about the history of Japanese mathematics, including Takakazu Seki's exploits in the Edo period, the progress of Japanese mathematics research after the Meiji period, when there was a full-scale introduction of western mathematics, and current problems resulting from the popularization of computers. Then, he described the relationship between mathematics and industrial technologies using the examples of cryptography, nanotechnology, and computer graphics. He suggested that

Photo 7. Prof. Masato Wakayama, Director, Institute of Mathematics for Industry, Kyushu University, giving an invited talk.

Fig. 1. Web site of NTT CS Labs Open House 2013.

mathematicians should discuss and share problems with industrial communities to improve both mathematics and industrial technologies.

6. Information transmission using the web

We have made continuous efforts to inform a large number of people about our research activities and results at NTT CS Labs. We set up both Japanese and English websites [1], [2] for Open House 2013 to improve the international dissemination of information (**Fig. 1**).

On the websites, we posted the open house program, venue, and access information, and put up abstracts of the research talks and exhibits in advance to enable visitors to obtain information about the event. After closing the event, the booklet, exhibition posters, and reference information were added to the websites (**Fig. 2**). Everyone can access this content. We will also upload videos of the director's keynote speech and the four research talks.

Fig. 2. Web page of an exhibit.

7. Concluding remarks

Just as they did last year, many visitors came to the NTT CS Labs Open House 2013 and engaged in lively discussions on the research talks and exhibits and provided many insightful opinions on the presented results. In closing, we would like to offer our sincere thanks to all of the visitors and participants who attended this event.

References

[1] Open House website (in Japanese).
http://www.kecl.ntt.co.jp/openhouse/2013/
[2] Open House website (in English).
http://www.kecl.ntt.co.jp/openhouse/2013/index_en.html

Akinori Fujino
Senior Research Scientist, Learning and Intelligent Systems Research Group, NTT Communication Science Laboratories.
He received the B.E. and M.E. degrees in precision engineering and the Ph.D. degree in informatics from Kyoto University in 1995, 1997, and 2009, respectively. He joined NTT in 1997. His current research interests include machine learning and information extraction from complex data.

Toshitaka Kimura
Senior Research Scientist, Sensory and Motor Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories.
He received the Ph.D. degree in human neuroscience from the Graduate School of Arts and Sciences, the University of Tokyo in 2003. He joined NTT in 2003. He has recently been engaged in research on sensorimotor control and action-perception interaction in humans. He is a member of the Society for Neuroscience, the Japan Neuroscience Society, and the Institute of Electronics, Information and Communication Engineers.

Kazushi Maruya
Senior Research Scientist, Sensory Representation Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories.
He received the M.A. and Ph.D. degrees in psychology from the University of Tokyo in 2001 and 2005, respectively. He joined NTT CS Labs in 2008. He was a researcher at the Intelligent Modeling Laboratory, University of Tokyo, from 2004 to 2005, and a visiting scientist at the Department of Psychology, Vanderbilt University, Nashville, TN, USA, from 2006 to 2008. His current research interests include visual motion processing for perception, especially for natural scene perception. He is also interested in human computer interface design for reading digital text.

Marc Delcroix
Research Scientist, Signal Processing Research Group, NTT Communication Science Laboratories.
He received the M.E. degree in engineering from the Free University of Brussels, Belgium, and the Ecole Centrale Paris, France, in 2003 and the Ph.D. degree from the Graduate School of Information Science and Technology, Hokkaido University in 2007. He is currently engaged in research on speech and audio signal processing.

Hiroaki Sugiyama
Researcher, Communication Environment Research Group, NTT Communication Science Laboratories.
He received the B.E. and M.E. degrees in mechano-informatics from the University of Tokyo in 2007 and 2009, respectively. He joined NTT in 2009. His current research interests include conversational dialogue systems and infant language development.

Papers Published in Technical Journals and Conference Proceedings

Power Reduction by Adaptively Optimizing Optical Power Using Actual BER for 10G-EPON Systems

N. Ikeda, H. Uzawa, K. Terada, S. Shigematsu, H. Koizumi, and M. Urano

Proc. of the 39th European Conference and Exhibition on Optical Communication (ECOC 2013), Vol. 2013, pp. 6.12, London, UK.

The optical line terminal calculates the pre-FEC BER using the number of corrected error bits and decides the optical power of the optical network unit (ONU) transmitter. The ONU adaptively adjusts the optical power according to the decision during the discovery window. The power consumption is reduced by 250 mW without any additional devices and without degrading throughput.

Interaction between Agency and Body-ownership in Terms of Schizophrenia and Schizotypy

T. Asai

Proc. of Tokyo Conference on Philosophy of Psychiatry 2013, Vol. 1, No. 1, p. 36, the University of Tokyo Komaba Campus, Tokyo, Japan.

Awareness of our own body (sense of body ownership) and action (sense of agency) is a fundamental component of self-consciousness. These sensory representations of the self are also important when we consider aberrant experiences such as delusions and hallucinations in patients with schizophrenia and also in the general population with schizotypal personality traits. I first introduce some empirical evidence that suggests their atypical representation in the sense of agency and body-ownership, respectively. On the other hand, these components of minimal self are closely related to each other and are integrated to form one agent with a unified awareness of the body and action. I propose that schizophrenia should not be regarded merely as a disorder of agency, but as a disorder of the hierarchic representation of the self where the sense of action, body, memory, and also identity must be integrated.

Wide-bandwidth Charge Sensitivity with a Radio-frequency Field-effect Transistor

K. Nishiguchi, H. Yamaguchi, A. Fujiwara, H. S. J. van der Zant, and G. A. Steele

Appl. Phys. Lett., Vol. 103, No. 143102, 2013.

We demonstrate high-speed charge detection at room temperature with single-electron resolution by using a radio-frequency field-effect transistor (RF-FET). The RF-FET combines a nanometer-scale silicon FET with an impedance-matching circuit composed of an inductor and capacitor. Driving the RF-FET with a carrier signal at its resonance frequency enables small signals at the transistor's gate to modulate the impedance of the resonant circuit, which is monitored at high speed using the reflected signal. The RF-FET driven by high-power carrier signals enables a charge sensitivity of 2×10^{-4} e/Hz$^{0.5}$ at a readout bandwidth of 20 MHz.

Performance Evaluation of Short-range MIMO Using a Method for Controlling Phase Difference between Each Propagation Channel

K. Sakamoto, K. Hiraga, T. Seki, T. Nakagawa, and K. Uehara

IEICE Trans. on Communications, Vol. E96-B, No. 10, pp. 2513–2520, 2013.

A simple decoding method for short-range multiple-input multiple-output (SR-MIMO) transmission can reduce the power consumption for MIMO decoding, but the distance between the transceivers requires millimeter-order accuracy in order to satisfy the required transmission quality. In this paper, we propose a phase difference control method between each propagation channel to alleviate the requirements for the transmission distance accuracy. In the proposed method, the phase difference between each propagation channel is controlled by changing the transmission (or received) power ratio of each element of sub-array antennas. In millimeter-wave broadband transmission simulation, we clarified that when sub-array antenna spacing is set to 6.6 mm and element spacing of the sub-array antenna is set to 2.48 mm, the proposed method can extend the transmission distance range satisfying the required transmission quality, which is a bit error rate (BER) before error correction of less than 10–2 from 9–29 mm to 0–50 mm in QPSK, from 15–19 mm to 0–30 mm in 16 QAM, and from only 15 mm to 4–22 mm in 64 QAM.

Hydrogen-enhanced Lattice Defect Formation and Hydrogen Embrittlement of Cyclically Prestressed Tempered Martensitic Steel

T. Doshida, M. Nakamura, H. Saito, T. Sawada, and K. Takai

Acta Materialia, Elsevier, Vol. 61, No. 1, pp. 7755–7766, 2013.

The number of lattice defects formed by applying cyclic prestress with/without hydrogen for various cycles and strain rates during cyclic prestress was compared for tempered martensitic steel. A tensile test was also carried out to evaluate hydrogen embrittlement susceptibility following the application of cyclic prestress. The results showed that when cyclic prestress was applied without hydrogen, the number of cycles and the strain rate had no apparent effect on mechanical properties or fracture morphology at the time of the subsequent tensile test. In contrast, when cyclic prestress was applied with hydrogen, the fracture strain and fracture stress decreased with an increasing number of prestress cycles and a decreasing strain rate, and the fracture morphology exhibited brittle fracture, signifying an increase in hydrogen embrittlement susceptibility at the time of the tensile test. The number of hydrogen-enhanced lattice defects also increased with an increasing number of cycles, and a decreasing strain rate was found when cyclic prestress was applied with hydrogen. These results indicate a correlation between hydrogen embrittlement susceptibility and the number of hydrogen-enhanced lattice defects. The increased hydrogen-enhanced lattice defects were probably vacancies and vacancy clusters formed by the interactions between hydrogen and dislocation movement during the application of cyclic prestress. The vacancies and vacancy clusters formed during the application of cyclic prestress with hydrogen presumably caused intergranular fracture and increased hydrogen embrittlement susceptibility.

Case Study of Model Adaptation: Transfer Learning and Online Learning

K. Imamura

Proc. of International Joint Conference on Natural Language Processing, pp. 1292–1298, Nagoya, Japan, 2013.

Many NLP tools are released as programs that include statistical models. Unfortunately, the models do not always match the documents that the tool users are interested in, which forces the user to update the models. In this paper, we investigate model adaptation under the condition that users cannot access the data used in creating the original model. Transfer learning and online learning are investigated as adaptation strategies. We test them on the category classification of Japanese newspaper articles. Experiments show that both transfer and online learning can appropriately adapt the original model if the dataset for adaptation contains all data, not just the data that cannot be well handled by the original model. In contrast, we confirmed that the adaptation fails if the dataset contains only erroneous data as indicated by the original model.

Highly Realistic 3D Display System for Space Composition Telecommunication

M. Date, H. Takada, S. Ozawa, S. Mieda, and A. Kojima

Proc. of IEEE IAS Annual Meeting, Vol. 2013-ILDC, No. 440, pp. 1–6, Orland, FL, USA, 2013.

We describe a highly realistic 3D display system that generates a composition of local and remote locations for telecommunication purposes. It uses a 3D projector and head tracking to display a person in a remote location as a life-size stereoscopic image against background scenery. Since it generates displayed images that correspond to the observer's viewing position, it well reproduces the fidelity of existence and the feel of a material. We also describe a simple, fast, and high quality algorithm for background scenery generation, the development of which was inspired by the visual effects of DFD (depth-fused 3D) displays. Our system is a promising means of achieving real-time communication between two different locations in cases where a sense of reality is required.

Colorless Optical Add/drop Using Small Matrix Switch and Cyclic AWG

T. Watanabe, S. Sohma, and S. Kamei

Proc. of the 18th Microoptics Conference (MOC'13), Vol. 1, No. 1, F-1, Tokyo, Japan, 2013.

We describe a new wavelength routing switch architecture that uses small matrix switches and cyclic arrayed-waveguide gratings. This switch enables us to provide colorless add/drop ports in reconfigurable optical add/drop multiplexer nodes.
