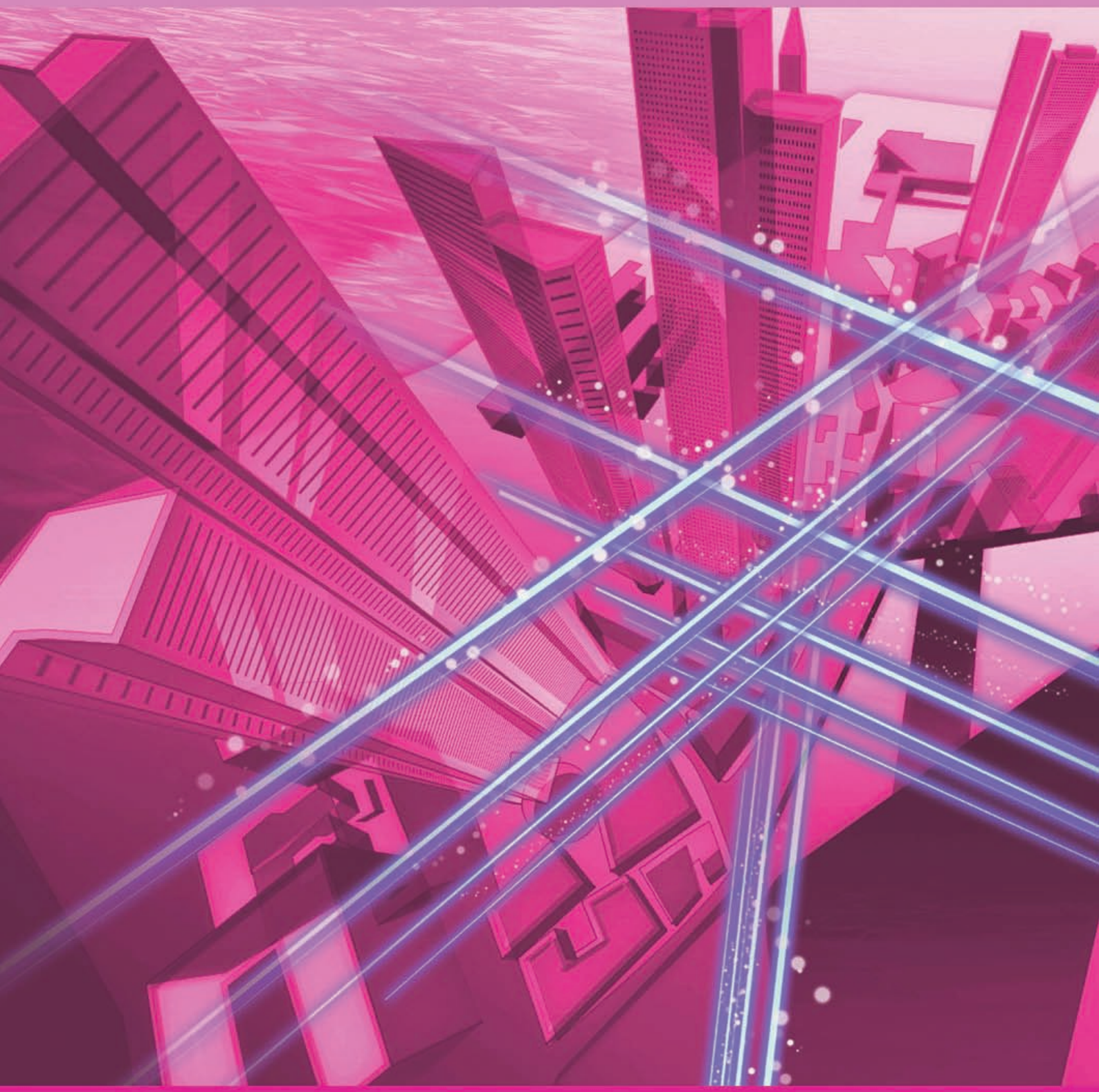
NTT Technical Review

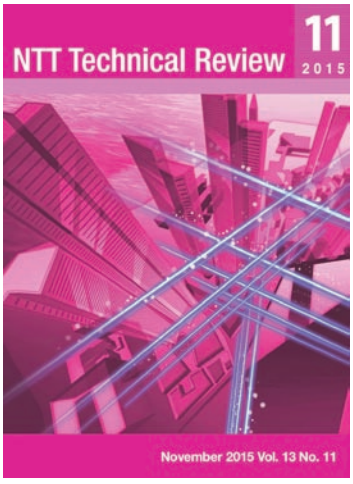
11
2015



November 2015 Vol. 13 No. 11

NTT Technical Review

November 2015 Vol. 13 No. 11



Comprehensive "Puzzle Solving" Based on Simple Ideas in the Age of Media Information Overflow

Kunio Kashino
Senior Distinguished Researcher,
NTT Communication Science Laboratories

As the volume of music, photographs, and video on the Internet continues to increase, the need for accurate and high-speed searching of media information is growing rapidly. We asked Dr. Kunio Kashino, Senior Distinguished Researcher at NTT Communication Science Laboratories, to tell us about the current state of research on media search in today's society and his thoughts on how researchers should view and approach their work.

Keywords: media search, media dictionary, matching

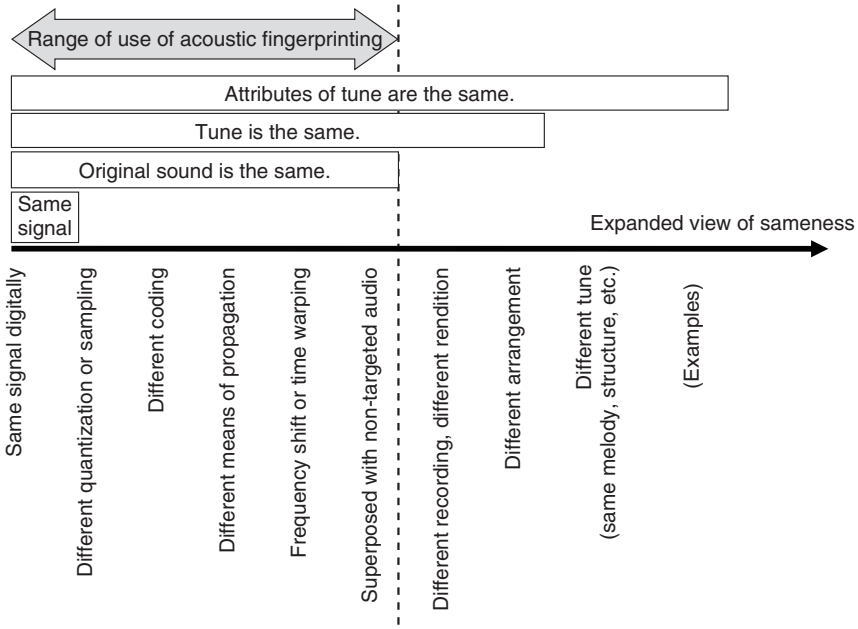
Aiming to create the first "media dictionary"

—Dr. Kashino, please tell us about your current research endeavors.

We call the physical means of conveying information in the form of sound, pictures, or video "media," and I am researching techniques for analyzing such means of conveying information and identifying its content by computer. I am working in particular on deciphering the information conveyed by some sort of medium by creating and referencing a media dictionary.

We can compare the referencing of a media dictionary with the analysis of text. For example, when we come across a word that we don't understand while reading something, we can look up that word in a dictionary. Furthermore, when analyzing character strings, we obtain a good result when we compare substrings with entries in a dictionary and find an item that matches. Here, the accuracy of analyzing character strings can be improved by preparing a dictionary with an extensive collection of entries. Similarly, we consider that preparing and referencing a media dictionary that assembles as many audio, picture, and video entries as possible will improve the accuracy of analysis. However, preparing a dictionary—even a conventional word dictionary—is not a trivial task, and in the case of media, even more difficult problems arise. This is why research in this area is needed.

One difficult problem is determining how to judge what is the same and what is different. In the case of words, a character sequence is the key to determining whether a match exists between two items. In the case of media, however, determining whether something is the same as something else is not that simple. For example, we can imagine the same song sung by different singers or the same tune with different arrangements, which means that we can treat two things as



There are various viewpoints on *sameness* in media data. Among these, identifying original signals as the same data has come to be called acoustic fingerprinting and video fingerprinting technology.

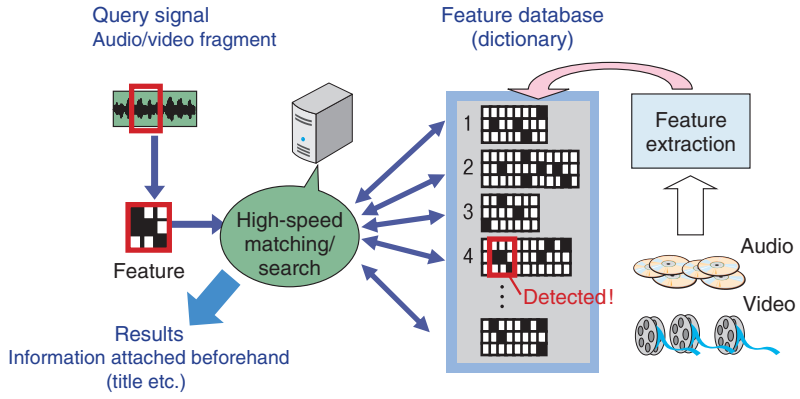
Fig. 1. Range of "same" media data (using music as an example).

being the same in one sense and as being different in another. Furthermore, for a person who does not regularly listen to music in a certain genre, all tunes in that genre may sound the same. In contrast, a person who is very familiar with music of that genre may be able to recognize a certain recording and tell which parts of that recording are especially moving. To that person, such fine differences may be very significant.

These are important issues, but in our research, we began by searching for media data that include audio or video that "sounds the same" or "looks the same" as a fragment of audio or video used as input. This type of technology eventually came to be called acoustic fingerprinting and video fingerprinting (**Fig. 1**). Even by narrowing down the problem in this way, media data itself can undergo major changes for a variety of reasons, such as sound being superposed with other sound at high volume or video being post-processed, so performing accurate searches for target media data is not that easy.

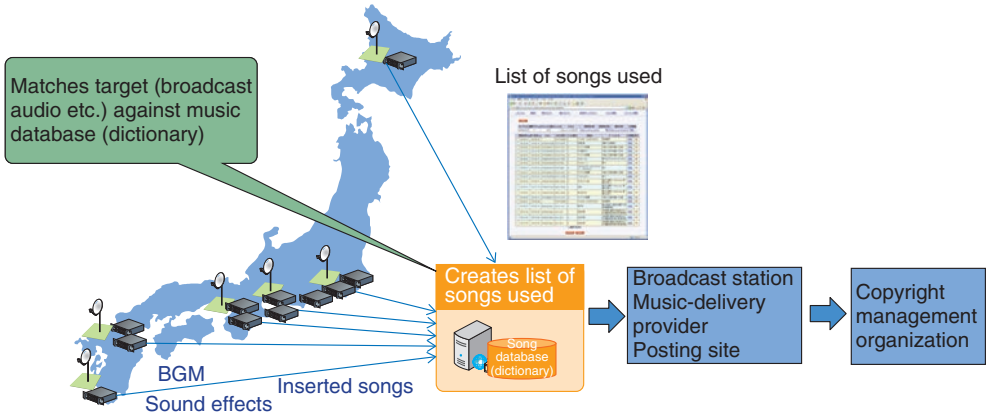
Now, if we assume that the sameness of media data can be determined by some method, there is still another problem that we must deal with, namely, search speed. Nowadays, individuals can easily create media data; on top of that, there is an increasing amount of media data being automatically generated. To search such media data, it has become a common practice to use metadata attached to the main body of data, that is, auxiliary information related to content. However, the effectiveness of using only external metadata attached without human intervention has its limits, and the amount of data is increasing so dramatically that it would take too long to process such data by manual means. There is therefore a need for technology that can automatically analyze and search the actual content of media data in an efficient manner (**Fig. 2**).

This dictionary referencing of media has already begun to be used in areas familiar to all of us. For example, copyright management of music and sound effects used in TV and radio programs is accomplished through such dictionary referencing (**Fig. 3**). More than five million sound sources are registered in the dictionary currently in use. In this method, a portion of each entry in the dictionary is compared with the actual broadcast audio from moment to moment, which makes it possible to instantaneously identify which sound source is being played at what hour, minute, and second on what channel. The basis for this technology was born about 20 years ago, but today, the speed of processing far exceeds what was



Processing is broadly divided into *feature extraction* and *high-speed matching*. Feature data are extracted from audio and video and stored beforehand. These data are then compared with partial features of the query signal at high speed. Quantification of features and high-speed search methods are current research topics.

Fig. 2. Mechanism for identifying "same" audio or video.



An NTT Group company has been providing a song-list creation service for music copyright management since 2007. This service can identify and list all songs used in target broadcasts and music-delivery services and can detect the use of songs in background music (BGM) and short-time use of music.

Fig. 3. Copyright management in song use.

possible at that time; in fact, it's several thousand times faster. In addition, it's not uncommon for multiple sounds to be superposed in the middle of a broadcast program, and good progress is being made on technology that can identify audio under such conditions without any problems.

—What idea led you to begin your research?

As a graduate student, I was interested in developing a way to recognize various types of sounds and noises, which is a very difficult problem. In any case, it is exceptionally hard to recognize individual sounds from a mixture of sounds by computer. However, the overlapping of sounds is not normally a problem to human beings in everyday life; indeed, it can be rather enjoyable, as in the case of musical chords. Furthermore, when listening to lyrics together with some sort of accompaniment, people can even make out phonemes that scarcely appear on the waveform, as in the case of unvoiced consonants. I found this to be a fascinating puzzle, as to what kind of mechanism

made this possible. After giving this much thought, I concluded that the mechanism would have to be some sort of matching, as done when checking a dictionary.

This is a very simple concept. The general perception of matching has been that it cannot be useful for processing complex information, or that it is not a viable approach in research. Regardless, I moved forward in my research one step at a time, having various thoughts such as "What would happen if all sounds and things in the world were described in a dictionary" and "Couldn't entries in such a dictionary be derived from actual data?" My idea of creating a media dictionary began around this time.

As an experiment, I tried to record the sounds of traffic during my commute to work. This recording included a variety of sounds such as wind and automobile noise. Since matching can be a robust process in the case of overlapping sounds, I thought that creating a dictionary of traffic sounds would enable me to analyze the soundscape in much the same way that text can be analyzed by referencing an ordinary dictionary.

Composite and comprehensive approach based on simple ideas

—Has your research progressed smoothly? Have you had any difficult experiences?

In 1998, more than two years after I began my research, I learned by chance that there was a need in the world for technology that could identify the appearance of previously registered audio and video content. Specifically, there was a need for detecting TV and radio commercials. At that time, the checking of broadcast commercials was carried out for the purpose of verifying the broadcast of certain commercials or as part of marketing surveys, but this was accomplished by visual means using human labor. However, it was clear that the technique that my colleagues and I were studying could do this checking much faster and more accurately than people could do manually. I lost no time in turning the core part of this technology into a software library and even wrote up an extensive programming guide book so that other people could make use of this technology. Much to my surprise, survey companies that target commercials came to adopt this technology one after another. In this way, I experienced how a simple idea could create value.

At the same time, there were many difficult things to deal with. The target applications naturally expanded, and the types and scale of dictionaries increased rapidly. In the early 2000s, we took up the problem of finding a way to send a song title by email to mobile phone users who would have their phones "listen" to some music that was being played out loud. To do this, we registered sound sources in a dictionary on a scale of several hundred thousand tunes to begin with. At this scale, we noticed interesting phenomena involving errors in recognition, which appeared similar to how two unrelated people may look similar by chance. Then, on increasing this scale by an order of magnitude, we encountered still other types of phenomena. Next, in 2008, we tried identifying known content, such as movies or music pieces, included in the movies posted on the Internet, but to target a scale corresponding to all posted movies, which were increasing on a daily basis, we saw that there would be no other way but to increase the processing speed by about 100 times. At first, I didn't think this was feasible, but on brainstorming with my colleagues, we came up with some ideas and somehow overcame this problem.

—Now that you have overcome some difficulties and achieved some results, are you close to achieving your objectives?

No, not yet! My awareness of the problem when I began my research 20 years ago was centered on the ingenious mechanisms that human beings use to hear sounds and see things. Since then, processing that, in a sense, far exceeds human capabilities has come to be realized, but I cannot say as yet that the original problem has been solved. Moreover, as I mentioned earlier, finding a way of determining what exactly is the same and what exactly is different is still an issue to be addressed. In recent years, it has become relatively easier to store and process huge amounts of media data, so we can consider that even newer approaches may be applicable to this problem.

Taking a puzzle-solving approach from diverse viewpoints

—From here on, how do you plan to approach your research?

I believe it is important that media data be analyzed in a composite and comprehensive manner. In daily information processing as performed by human beings, a person unconsciously mobilizes a variety of

senses starting with sight and hearing to understand the surrounding environment. From the beginning, I have had an interest in solving problems in a concrete manner. Real-world problems are often complex and appear in all sorts of forms. Starting with the manifestation of some kind of phenomenon, the task is to work backwards to identify and analyze the problem. Here, however, being able to solve the problem from only one viewpoint is quite rare; a researcher needs to think in a composite and comprehensive manner. This can be compared to the way in which a general physician treats the "whole" patient instead of just focusing on a particular organ.

I would also like to look at problems from a broader field of view, not just from the viewpoint of a researcher. To give an example of what I mean here, let me take you back to the time when it became possible for mobile phones to take videos. At that time, we performed a demonstration of how information obtained from a TV screen captured with a mobile phone could be used to take the user to a website that would provide related information on the program being shown. Technically speaking, the ability to identify screen content from a small and faint image was quite interesting at that time. However, our project was actually a failure. In actuality, viewers are not interested in pressing a button on their mobile phone and capturing video while watching TV. This is a perfect example of how researchers can become self-righteous in their application of technology.

From this experience, I realized that I would like to be motivated by "solving puzzles" in my research while taking to heart the need to approach things from a variety of standpoints and views.

—Dr. Kashino, can you leave us with some advice for young researchers?

I feel that establishing a research theme is very important. I want young researchers to pursue a theme that they believe to be important—not simply a theme that is currently trendy in society. However, at that time, you should not become self-righteous about the theme that you want to choose; you must examine closely why you feel that theme is important and whether anyone else thinks it to be important. If no one at all is working on that problem, perhaps it could be that it is simply not important. However, the best scenario is to find a problem that almost no one else thinks is important but is, in reality, extremely important. Such a research theme will likely grow and develop like the trunk of a tree. Additionally, when you find yourself up against an obstacle, it's a good policy to reconsider what, in the end, is really of importance in your research.

■ **Interviewee profile**
Kunio Kashino

Senior Distinguished Researcher and Head of Media Information Laboratory, NTT Communication Science Laboratories.

He received a Ph.D. from the University of Tokyo for his pioneering work on music scene analysis in 1995. Since joining NTT in 1995, he has been working on audio and video analysis, search, retrieval, and recognition algorithms and their implementation. He has received several awards including the Maejima Award in 2010, the Young Scientists' Prize for Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2007, and the IEEE (Institute of Electrical and Electronics Engineers) Transactions on Multimedia Paper Award in 2004. He is a senior member of IEEE. He is also a Visiting Professor at the National Institute of Informatics, Tokyo, and at the Graduate School of Information Science and Technology, the University of Tokyo.

Embracing Information Science and Technology—Decoding, Exploring, and Designing the World

Eisaku Maeda

Abstract

The era in which human beings are confronted with machines (computers or artificial intelligence) as disparate elements is coming to an end. From here on, we will embrace information science and technology as part of ourselves. This will necessitate the ability to decode, explore, and design the entire world, including us human beings. While bearing in mind the drastic changes in the information environment that we have experienced in the first fifteen years of the twenty-first century, we must think about what should make up the basic research that will form the compass of the future as we envision the year 2030, fifteen years from now.

Keywords: basic research, innovation, information science

1. Introduction

The recent era is one in which human beings have *faced off* against "machines" (computers or artificial intelligence), so to speak. This era is coming to an end. From this point on, we will embrace information science and technology as part of ourselves. This means that we will need to be able to decode, explore, and design everything in the entire world, including us human beings. Consequently, we must think about what the basic research will be that will form the compass of the future as we envision the year 2030, fifteen years from now, while still keeping in mind the drastic changes in the information environment that we have so far experienced in the twenty-first century. Thanks to the development of information science and technology, the real world in which we embrace information science and technology and are in turn embraced by it, is changing greatly in the following three ways.

1.1 From measuring to understanding

The first turning point is a transition from an era in which physical quantities in the real world are measured by sensors to an era in which diverse information flows through two types of space-time environments—the real world and the virtual world—and are decoded. Sound-recording microphones will be replaced by microchips that decode the audio environment. Such a development will be equivalent to the evolution of human beings' audiovisual processing system, which includes the sensory organs such as the ears and eyes and the frontal lobe of the cerebral cortex. The growing intelligence of sensor devices that continue to evolve in the real world can be said to be the origin of the aforementioned turning point. A similar evolution is happening in the virtual world. This development will necessitate new security technologies.

1.2 From analysis to exploration

Second, we will transition from an era of analysis, in which massive amounts of data are gathered and analyzed using statistical techniques, to an era of exploration, in which the conclusions necessary for control or decision-making can be immediately acquired.

There are two major features to exploration in the age of big data. The first is that an exploratory result will be presented with a probability value attached.

The second is the possibility of fast yet inexpensive exploration. This will be the key to the new age. With the arrival of big data, an empirical element is being added to the academic discipline that is information science. These two features can be likened to the key of improving productivity through assay (screening) in the experimental sciences and manufacturing science.

1.3 From implementation to design

The third aspect is the move from an age of implementation in which information processing technologies in the form of machines are actuated in the real world, to an age in which design is optimized for the overall system that connects both the real world and the virtual world into a cyber-physical system. Because the enterprise to design the overall world is itself a decoding application, recursive methods will expand our understanding of the world outward in a spiral. The research we are tackling right now can be placed in the flow of decoding, exploring, and designing the world.

2. Communication science as the compass for the future

The research topics being tackled by NTT Communication Science Laboratories today can be placed in the flow of decoding, exploring, and designing the world. If we classify the technologies introduced in these Feature Articles into these three categories, we can consider the technologies introduced in "Biological Measures that Reflect Auditory Perception" [1] and "Deep Learning Based Distant-talking Speech Processing in Real-world Sound Environments" [2] to be technologies for decoding the real world. We can consider the technology described in "Combinatorial Optimization Using Binary Decision Diagrams" [3] to be technology for searching in the real world, and those in the articles "Generative Modeling of Voice Fundamental Frequency Contours for Prosody Analysis, Synthesis, and Conversion" [4] and "Yu bi Yomu: A New Text Display System Using Tracing Behavior" [5] to be technologies for designing the real world. Another technology for designing the real world is described in "'Hen-Gen-Tou (Deformation Lamps)'—Amazing Illumination to Make Static Objects Dynamic" [6].

The information environment that permeates our lives has dramatically changed since the beginning of the twenty-first century, and the intensity of the transformation is continuing to grow. Even in the field of basic research, we are now in an era in which we must choose the challenges to tackle and we must contribute to introducing new technologies to the market with a sense of urgency of the times. As NTT seeks to create new markets by exploiting *Co-Innovation* through collaboration with companies in other industries, it can be said that, in fact, the expectation placed on the value of basic research and its fruits is becoming even greater. Each of the fruits born from basic research is a valuable seed of innovation. They are waiting for the arrival of the right time to blossom [7, 8].

Just when and where are technologies with the potential for innovation born? Understanding and exploiting this insight quickly is the key to winning the competition to create services, even in the field of research and development. The mission of private basic research is to diligently refine technologies that blossom a few years or a decade from now as we respond to the demands of the times. It is also to resolutely continue the intellectual challenge of creating a new world that others have not arrived at yet by creating new knowledge and patiently verifying it to build it up [9]. The research achievements of NTT Communication Science Laboratories are posted on our website as the occasion arises. We also introduce our efforts at the open house held in June every year [10].

References

[1] S. Furukawa, S. Yamagishi, H-I Liao, M. Yoneya, S. Otsuka, and M. Kashino, "Biological Measures that Reflect Auditory Perception," NTT Technical Review, Vol. 13, No. 11, 2015. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201511fa3.html

[2] S. Araki, M. Fujimoto, T. Yoshioka, M. Delcroix, M. Espi, and T. Nakatani, "Deep Learning Based Distant-talking Speech Processing in Real-world Sound Environments," NTT Technical Review, Vol. 13, No. 11, 2015. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201511fa4.html

[3] M. Nishino, N. Yasuda, T. Hirao, S. Minato, and M. Nagata, "Combinatorial Optimization Using Binary Decision Diagrams," NTT Technical Review, Vol. 13, No. 11, 2015. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201511fa6.html

[4] H. Kameoka, "Generative Modeling of Voice Fundamental Frequency Contours for Prosody Analysis, Synthesis, and Conversion," NTT Technical Review, Vol. 13, No. 11, 2015. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201511fa2.html

[5] K. Maruya and J. Watanabe, "Yu bi Yomu: A New Text Display System Using Tracing Behavior," NTT Technical Review, Vol. 13, No. 11, 2015. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201511fa5.html

[6] NTT press release issued on Feb. 17, 2015. http://www.ntt.co.jp/news2015/1502e/150217a.html

[7] E. Maeda, “The Evolution of Basic Research,” NTT Technical Review, Vol. 12, No. 11, 2014.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201411fa1.html
[8] E. Maeda, “Basic research—Defining our age and the future - The origin of ideas and the seeds of innovation,” Director’s address of NTT Communication Science Laboratories Open House 2014, Kyoto, Japan, June 2014.
http://www.kecl.ntt.co.jp/openhouse/2014/talk/director/index_en.html
[9] E. Maeda, “The Enchantment of Turning toward Academics,” IEICE Information and System Society Journal, Vol. 19, No. 2, pp. 21–22, 2014 (in Japanese).
[10] Website of NTT Communication Science Laboratories Open House 2015.
http://www.kecl.ntt.co.jp/openhouse/2015/index_en.html

Eisaku Maeda
Vice President, Head of NTT Communication Science Laboratories.
He received a B.E. and M.E. in biological science and a Ph.D. in mathematical engineering from the University of Tokyo in 1984, 1986, and 1993. He joined NTT in 1986. He was a guest researcher at the University of Cambridge, UK, in 1996–1997. He was awarded IPSJ’s 45th anniversary best paper on the next 50 years of information science and technology for his paper “Resurgence of Fairies and Goblins—A Proposal for the New Vision of “Ambient Intelligence.” His research interests are statistical machine learning, intelligence integration, and bioinformatics. He is a fellow of IEICE (Institute of Electronics, Information and Communication Engineers of Japan) and a senior member of IEEE (Institute of Electrical and Electronics Engineers) and IPSJ (Information Processing Society of Japan).

Generative Modeling of Voice Fundamental Frequency Contours for Prosody Analysis, Synthesis, and Conversion

Hirokazu Kameoka

Abstract

This article introduces a state-of-the-art technique that makes it possible to convert speech to different speaking styles through the manipulation of the fundamental frequency (F_0) contour without destroying the naturalness of the speech. This technique can be used, for instance, to convert non-native speech to native-like speech, and to convert normal speech to speech with a more lively intonation similar to the way broadcasters speak. It can also be incorporated into text-to-speech systems to improve the naturalness of computer-generated speech.

Keywords: prosody, intonation, accent, voice fundamental frequency contour, generative model

1. Introduction

The fundamental frequency (F_0) contour refers to the time course of the frequency of the vocal fold vibration of speech. We usually use not only words and sentences to convey messages to the listener in daily communication, but also F_0 contours to add extra *flavor* to speech such as the identity, intention, attitude, and mood of the speaker. It is also important to note that the naturalness of F_0 contours is one of the most significant factors that affect the perceived naturalness of speech as a whole. In fact, synthesized (artificially created) speech with an unnatural F_0 contour often sounds robotic, lifeless, or emotionless. This article introduces a technique that makes it possible to convert the speaking style of an input utterance into different speaking styles by controlling the F_0 contour while retaining its naturalness.

The proposed technique can be used, for example, to convert the intonation of the speech uttered by a non-native speaker to a more fluent intonation similar to the way native speakers speak, to convert the accents of speech to those with different dialects, and to convert the intonation of normal speech to a more lively intonation similar to the way broadcasters speak. It would also allow us to modify the intonation or accents of the *acted* speech by actors or actresses as desired without the need to retake the scene. It can also be incorporated into text-to-speech (TTS) systems to improve the naturalness of computer-generated speech. Furthermore, we can build a self-training system to assist students in improving their presentation and language skills. We are also interested in applying the proposed technique to develop a speaking-aid system that makes it possible to convert electrolaryngeal speech to normal speech, which can be used to assist people with vocal disabilities.

2. Fundamental frequency contour (intonation and accent)

The F_0 contour of speech consists of intonation and accent components. The intonation component corresponds to the relatively slow F_0 variation over the duration of a prosodic unit, and the accent component corresponds to the relatively fast F_0 variation in an

accented syllable. Both of these components are characterized by a fast rise followed by a slower fall. The former usually contributes to phrasing, while the latter contributes to accentuation during an utterance. In Japanese, for example, changing the positions of accents results in speech with different dialects or meanings. The magnitudes of these components correspond to how much emphasis the speaker intends to place on the associated phrase or accent. Thus, the magnitudes and positions of these components assist the listener in interpreting an utterance and draw attention to specific words. The F_0 contour also plays an important role in conveying to the listener various types of non-linguistic information such as the identity, intention, attitude, and mood of the speaker.

3. Generative modeling of F_0 contours

3.1 F_0 control mechanism

F_0 contours are controlled by the thyroid cartilage, which sits in front of the larynx. The assumption that the F_0 contour of speech consists of intonation and accent components is justified by the fact that the thyroid cartilage involves two mutually independent types of movement with different muscular reaction times. Specifically, the intonation and accent components respectively correspond to contributions associated with the translation and rotation movements of the thyroid cartilage. In the late 1960s, Fujisaki proposed a well-founded mathematical model that describes an F_0 contour as the sum of these two contributions [1, 2] (**Fig. 1**). This model approximates actual F_0 contours of speech fairly well when the model parameters are appropriately chosen, and its validity has been demonstrated for many typologically diverse languages. If we can estimate the movements of the thyroid cartilage automatically from a raw F_0 contour, we can simulate or predict the F_0 contour that we may observe when the thyroid cartilage moves differently, by simply modifying the values of the motion parameters. Since the movements of the thyroid cartilage are characterized by the levels and timings of the intonation and accent components, one important challenge is to solve the inverse problem of estimating these components directly from speech. However, this problem has proved difficult to solve. This is because it is difficult to determine a unique intonation and accent component pair only from their mixture (namely an F_0 contour), in the same way that it is impossible to determine a unique X and Y pair only from X + Y = 10. Several techniques have already been developed but so far with limited success.

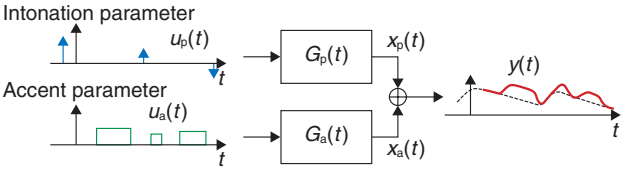


Fig. 1. F_0 control model (Fujisaki model).

3.2 Statistical formulation

There are some clues that can possibly help solve this problem. That is, we can use the fact that the levels and timings of intonation and accent components are statistically biased in normal speech. The author has proposed constructing a stochastic counterpart of the Fujisaki model that makes it possible to use statistical inference techniques to accurately and efficiently estimate the underlying parameters of the Fujisaki model [3, 4]. The problem of estimating the intonation and accent components from a raw F_0 contour is somewhat similar to the audio source separation problem. Audio source separation refers to the problem of separating the underlying source signals from mixed signals. Even though it looks as difficult as solving the X + Y = 10 problem, a statistical approach utilizing the statistical properties and the statistical distribution of the waveforms of audio signals has proved effective in solving it. The idea for the proposed method was inspired by this idea (**Fig. 2**).

An example of the estimated parameters related to the intonation and accent components (blue and green lines) along with the estimated F_0 contours (red line) plotted on the spectrogram of the input speech is shown in **Fig. 3(a)**. An example of converted speech with magnified accent components is shown in **Fig. 3(b)**, and an example of converted speech with shifted positions of accent components is in **Fig. 3(c)**. It is important to note that in these examples we were able to convert the speaking style of the input speech into different styles (one with prominent accents and the other with a different dialect) while retaining the naturalness as if the same speaker were uttering the same sentence in a different way. This was made possible because the proposed framework allows us to modify F_0 contours in such a way as never to violate the physical constraint of the actual F_0 control mechanism.

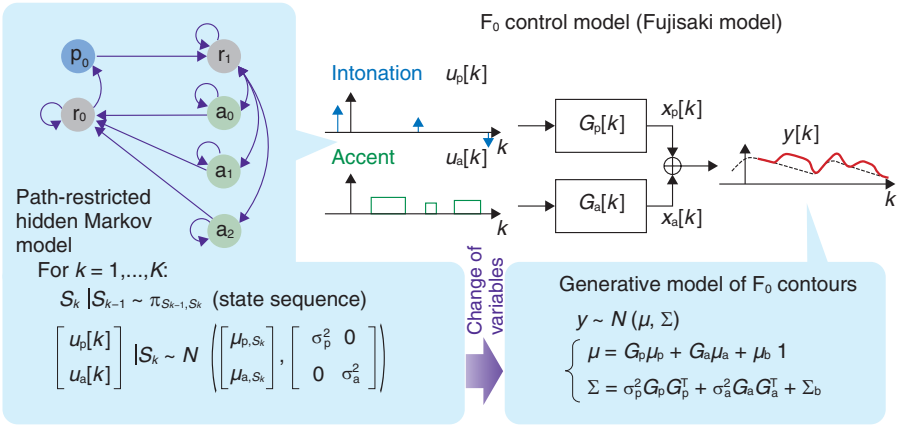


Fig. 2. Proposed model (a stochastic counterpart of the Fujisaki model, described by a discrete-time stochastic process).

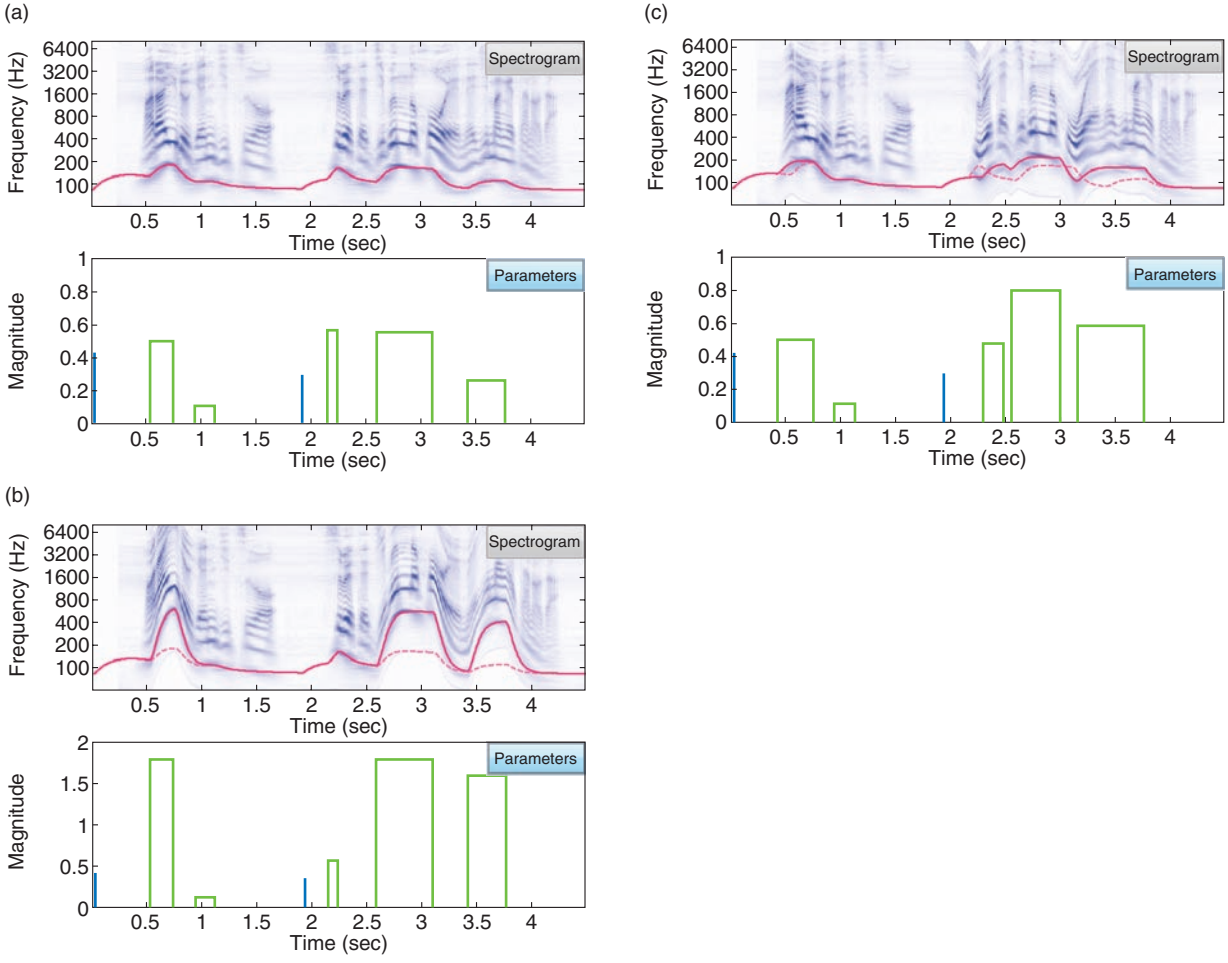


Fig. 3. (a) Example of the estimated parameters related to intonation and accent components (blue and green lines) along with the estimated F_0 contours (red line) plotted on the spectrogram of the input speech; (b) example of converted speech with magnified accent components; (c) example of converted speech with shifted positions of accent components.

4. Future perspective

While linear predictive coding (LPC), proposed in the late 1960s, led to the development of a speech analysis/synthesis system that provides a powerful parameter estimation framework for the vocal tract model, the proposed technique enables a new system that does the same for the F_0 control model (i.e., the Fujisaki model). In a way similar to the development of LPC, which has given rise to modern speech analysis/synthesis and become the cornerstone module for today's mobile and voice-over-IP (Internet protocol) communication, this work can also potentially open the door to a brand new speech analysis/synthesis framework.

Acknowledgment

This work was carried out in collaboration with the members of Sagayama/Moriya/Kameoka laboratory of the University of Tokyo. I thank all the people who contributed to this work. This work was supported by JSPS KAKENHI Grant Number 26730100, 26280060.

References

[1] H. Fujisaki and S. Nagashima, "A Model for the Synthesis of Pitch Contours of Connected Speech," Annual Report of the Engineering Research Institute, The University of Tokyo, Vol. 28, pp. 53–60, 1969.

[2] H. Fujisaki, "A Note on the Physiological and Physical Basis for the Phrase and Accent Components in the Voice Fundamental Frequency Contour," *Vocal Physiology: Voice Production, Mechanisms and Functions*, pp. 347–355, Raven Press, New York, USA, 1988.

[3] H. Kameoka, J. Le Roux, and Y. Ohishi, "A Statistical Model of Speech F_0 Contours," Proc. of SAPA 2010 (the 2010 Workshop on Statistical and Perceptual Audition), pp. 43–48, Makuhari, Japan, Sept. 2010.

[4] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino, "Generative Modeling of Voice Fundamental Frequency Contours," IEEE/ACM Trans. Audio, Speech and Language Processing, Vol. 23, No. 6, pp. 1042–1053, 2015.

Hirokazu Kameoka
Distinguished Researcher, NTT Communication Science Laboratories.
He received his B.E., M.S., and Ph.D. from the University of Tokyo in 2002, 2004, and 2007. His research interests include audio, speech, and music signal processing and machine learning. He received 13 awards over the past 10 years, including the IEEE (Institute of Electrical and Electronics Engineers) Signal Processing Society 2008 SPS Young Author Best Paper Award. He is the author or co-author of about 90 articles in journal papers and peer-reviewed conference proceedings. He is currently also an Adjunct Associate Professor at the University of Tokyo.

Biological Measures that Reflect Auditory Perception

Shigeto Furukawa, Shimpei Yamagishi, Hsin-I Liao, Makoto Yoneya, Sho Otsuka, and Makio Kashino

Abstract

Brain processes involved in auditory perception are reflected in various physical/physiological responses. Our recent studies indicate that in addition to brainwaves, responses that might seem to have nothing to do with audition—for example, pupillary responses, sounds emitted from the ear, and rhythmic finger-tapping movements—actually provide information about subjective auditory experiences of listeners. These findings may lead to various applications such as designing auditory displays with pleasing sounds adapted to individual listeners and developing novel techniques for diagnosing or compensating for impaired hearing.

Keywords: audition, biological measures, cognitive neuroscience

1. Introduction

Knowing how listeners perceive sounds or how sound information is processed in the brain is important not only in understanding basic mechanisms of the auditory system, but also in establishing guidelines for designing effective auditory displays or in evaluating devices for hearing aids. The use of psychological tests is a standard approach to examine perception, but it has limitations. There is no guarantee that the listener will be able to report his/her sensations accurately when the sensation does not have an associated *correct* answer or cannot be expressed in words or by pressing a button. It can also happen that the sensation varies each time even when the same stimuli are presented repeatedly to the same person. Furthermore, the results of a psychological test do not directly indicate what biological mechanisms underlie the sensation.

Biological measures are attracting increasing attention as an alternative approach to probing auditory perception. Recent technological advances in measurement techniques and accumulated knowledge in cognitive neuroscience have brought us new tools to examine how we hear. Biological measures allow us to gain information about the listener's perception and its underlying mechanisms without requiring the listeners to report their sensations explicitly.

2. Probing auditory perceptual content through neural activity

Recent attempts to probe perceptual content have mainly focused on cortical neuron activity. However, we should not forget that what can be observed as cortical neuron activity is only a small portion of the functions of the entire sensory system. We emphasize that the brainstem, often viewed as a physiological component of a low-level processing stage, plays an important role in perception. We are studying the brainstem by focusing on the frequency-following response (FFR). The FFR is a class of auditory evoked potentials, which can be recorded through electrodes placed on the scalp.

The FFR waveform is isomorphic to the stimulus waveform (**Fig. 1(c)**), and is believed to reflect characteristics of brainstem mechanisms for processing stimulus temporal structures. When tone bursts A and B, which differ in frequency, are presented alternatingly (ABA-ABA-ABA-…; **Fig. 1(a)**), a typical listener reports two patterns of percepts (**Fig. 1(b)**). One pattern is a single stream in which the short phrase

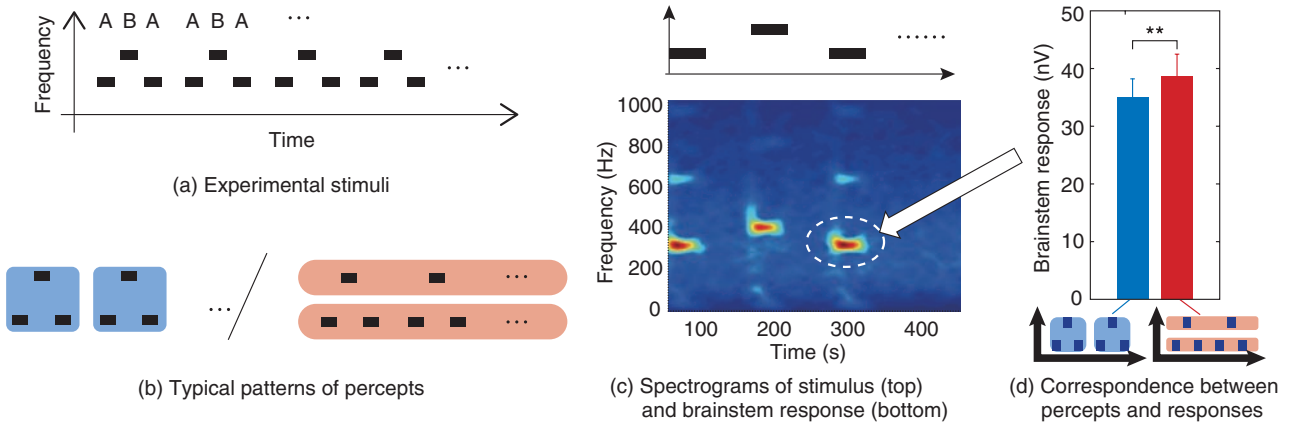


Fig. 1. Sound stimuli that evoke ambiguous percepts, and auditory brainstem responses.

ABA- is repeated. The other consists of two parallel streams with the same frequencies, that is, A-A-A-A- and -B---B---. The perception of those two patterns is not stable; the two patterns switch spontaneously and randomly during a prolonged listening session. Since the physical parameters of stimuli are unchanged, such perceptual switching should reflect what is happening in the individual listener's brain. Our research group discovered that the instantaneous perceived pattern and the strength of the FFR correlate with each other (**Fig. 1(d)**) [1].

Although the stimuli used in the experiment were simple and artificial, the discovery has marked implications in everyday listening. In natural environments, the sounds around us are not always clear. Nevertheless, the brain silently and continuously explores rational interpretations based on ambiguous auditory information. We consider that the spontaneous switching of the perception of the ABA sequence reflects brain processes involved in this exploration process. Our discovery indicates that supposedly low-level brainstem mechanisms play a significant role in interpreting acoustic information, which is usually regarded as a high-level process. This demonstrates an important contribution of the cross-level neural network in auditory perception. Our achievement also has a marked technological implication; it presents possibilities that continuously changing auditory perceptions within individual listeners can be captured as brainwaves.

3. *Reading* auditory perceptual content through the eyes

Electrical signals from neurons are not the only indicators reflecting brain activity. Our research group is also examining the eyes to study audition. The locus coeruleus, a brainstem nucleus, is known to contribute to controlling alertness and selective attention, and its neural activity appears to be reflected as pupil diameter [2]. We tested whether this mechanism could be used to evaluate the salience (tendency to attract attention) of sounds. We presented various sound samples, including environmental sounds and abstract tones, while recording the listener's pupil diameter. The results showed not only that the pupil dilates after the onset of a stimulus, but also that the strength of the dilation response correlates well with the subjective salience rating of the sounds (**Fig. 2**) [3]. The relationships between pupil diameter and cognitive load or attention have already been reported. However, to our knowledge, we were the first to find a relationship between pupil diameter and a basic subjective property of sound. Critical questions as to the acoustic properties of sound and the biological mechanisms that evoke pupil dilation response are yet to be studied. Nevertheless, it may become possible in the future to measure the degree to which listeners are attracted to given sounds by observing their eyes.

4. The ear's protection mechanism

The brainstem not only relays auditory signals from the ears to higher-level processing stages but also has

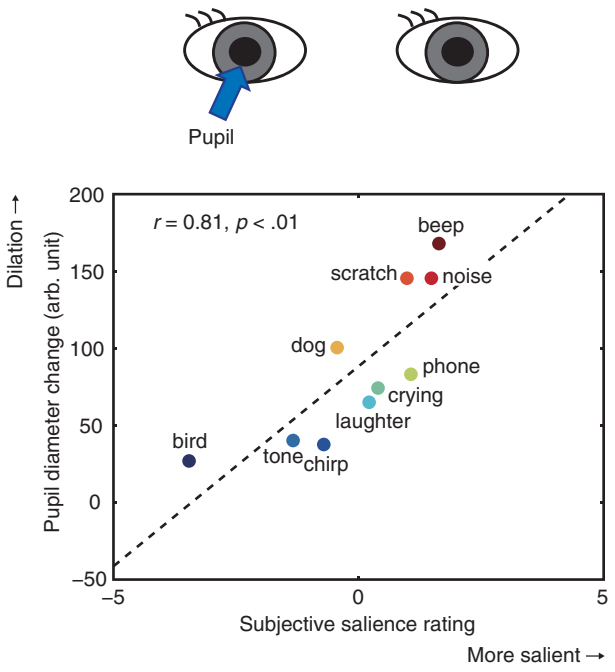


Fig. 2. Relationship between pupil diameter and subjective rating of sound salience.

downward pathways (referred to as the efferent system) that control the ears. Currently, however, there is as yet no standard view as to what practical role the efferent system plays in a real environment.

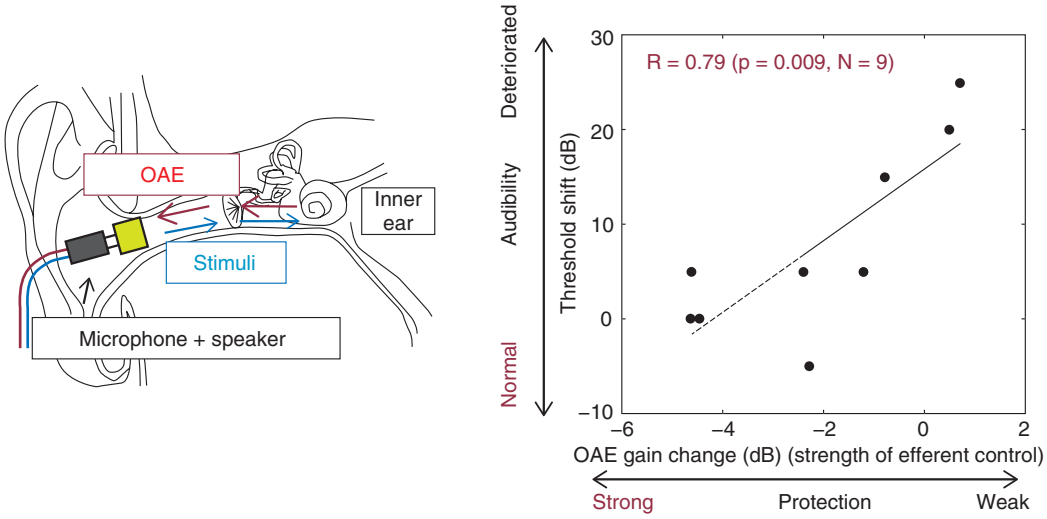
The inner ear does not simply transmit or convert acoustic signals but actively and mechanically amplifies the signals. This nonlinear amplifier can return the vibrations back to the eardrum, thereby emitting sounds (known as otoacoustic emissions or OAEs; **Fig. 3(a)**). The OAEs provide information about the amplification gain of the inner ear. It is known that the amplification gain often decreases when an intense sound is presented, which is likely due to the efferent system as mentioned earlier. This mechanism may function to protect the delicate inner ear from damage caused by intense sounds.

To test this hypothesis, we designed experiments involving violinists (violin students at a music school). In their daily instrument practice, violinists are exposed to intense sounds emitted from the instruments close to their ears. In fact, we observed significant, although temporary, elevations in their hearing threshold (known as a temporary threshold shift, TTS) after each practice session. The magnitude of TTS varied among individuals. We found a marked correlation between the TTS magnitude and the degree of the individual's amplification gain adjustment evaluated with OAE measurements (**Fig. 3(b)**) [4]. The results demonstrate that the efferent system serves as a protection mechanism. Noise-induced hearing loss is becoming a social issue in modern society, in which generations of people from teenagers to the elderly listen to music for many hours a day through the earphones of a portable audio device. In the future, evaluations of the functionality of the efferent system, as was done in the present study, may be used to estimate an individual listener's risk of noise-induced hearing loss, which would help to prevent hearing impairment.

5. Objective audiometry

Auditory detection thresholds in clinical audiometry are simple but very important measures in audiology since they are used as principal data for diagnosing hearing impairment. However, traditional methods for measuring detection thresholds have a critical problem. They are all subjective methods, which means that listeners report their percepts themselves. If the listener cannot report the percepts or deliberately reports them incorrectly, the methods are unreliable. Consequently, objective methods are sometimes used as an alternative. However, existing objective methods are used to evaluate the physiological functions of particular structures such as the inner ear and brainstem, but they do not directly indicate the listener's perceptions. Our research group has built on our accumulated knowledge and experience in sensory and motor studies in attempting to develop a conceptually novel objective method.

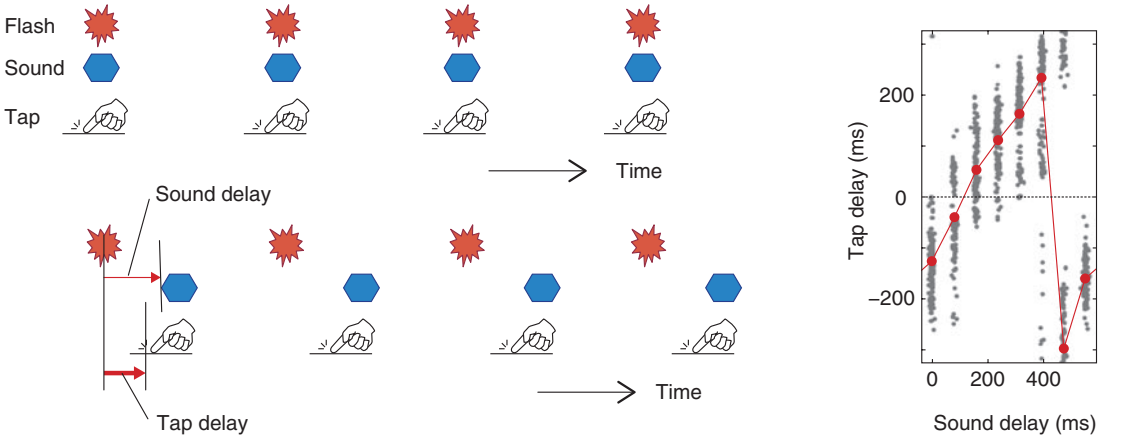
Our idea incorporates a so-called sensorimotor synchronization task in which the participant is required to tap their finger rhythmically and synchronously with iso-interval visual flashes. The subject performance in this task is known to be strongly influenced by the timing of simultaneously presented auditory stimuli (e.g., tone bursts). When the timing of the auditory stimuli is slightly delayed relative to that of the visual stimuli, the tap timing tends to shift to the auditory timing, even though the participant does not notice this (**Figs. 4(a)** and **4(b)**). Our study quantitatively indicated that this disturbance by auditory stimuli occurs even with very weak sounds that are close to the detection threshold [5]. This phenomenon is the principle of our objective method; that is, by examining the presence of the auditory disturbance, we can determine whether the sound is audible or not to the participant, without requiring subjective



(a) Measuring OAEs

(b) Evaluation of the protection function

Fig. 3. Evaluation of the protection function by measuring otoacoustic emissions OAEs.



(a) Disturbance by auditory stimuli

(b) Measurement results

Fig. 4. Synchronized tapping with rhythmic flashes.

reports. It should be noted that this method involves psychophysical tasks but is nearly *feigning-proof*. In this task, it is difficult for a participant to pretend that an audible sound is not audible, because the performance of the task would be *poorer* for *audible* sounds than for inaudible ones, unlike typical psychophysical tasks (including those adopted in traditional audiometry), in which the opposite would be true. As well as being a feigning-proof test, our new method is expected to be used as a diagnostic tool to reveal hearing problems that have not been sufficiently described by traditional methods.

6. Future perspectives

The findings described in the present article are exciting, as they open doors to new technologies to measure hearing. They also remind us that our auditory perception involves not only the sensory system but also complex interactions of multiple biological systems such as the motor, autonomic, and endocrine systems. New cognitive-neuroscientific paradigms

are demanded in order to understand and make use of such complex mechanisms. Rapidly developing sensing technologies and/or machine-learning techniques may provide solutions. We also should not forget the importance of conducting basic research that disentangles complex systems step-by-step. In this article, we presented new ideas incorporating concepts that would initially appear to be unrelated to audition. Those new ideas would not have been conceived without referring to the painstaking research efforts of many scientists in various fields.

Acknowledgment

The work described in sections 2 and 5 include products of the SCOPE project (121803022) commissioned by the Ministry of Internal Affairs and Communications. The study described in section 4 was performed in collaboration with Kyoto City University of Arts.

References

[1] S. Yamagishi, T. Ashihara, S. Otsuka, S. Furukawa, and M. Kashino, "Neural Correlates of Auditory Streaming in the Human Brainstem," Abstract of 36th Annual Midwinter Meeting of Association for Research in Otolaryngology (ARO Midwinter Meeting), Vol. 36, pp. 84–85, Baltimore, USA, Feb. 2013.

[2] G. Aston-Jones and J. D. Cohen, "An Integrative Theory of Locus Coeruleus-norepinephrine Function: Adaptive Gain and Optimal Performance," Annu. Rev. Neurosci., Vol. 28, pp. 403–450, 2005.

[3] H.-I. Liao, S. Kidani, M. Yoneya, M. Kashino, and S. Furukawa, "Correspondences Among Pupillary Dilation Response, Subjective Salience of Sounds, and Loudness," Psycho. Bull. & Rev., DOI 10.3758/s13423-015-0898-0, 2015.

[4] S. Otsuka, M. Tsuzaki, J. Sonoda, and S. Furukawa, "Effects of Short-duration Instrument Practice on the Auditory Peripheral Functions of Violin Players," Abstract of 38th ARO Midwinter Meeting, Vol. 38, p. 307, Baltimore, USA, Feb. 2015.

[5] S. Furukawa, K. Onikura, S. Kidani, M. Kato, and N. Kitagawa, "An Objective Measure of Auditory Detection Threshold based on a Light-synchronized Tapping Task," Abstract of 38th ARO Midwinter Meeting, Vol. 38, p. 157, Baltimore, USA, Feb. 2015.

Shigeto Furukawa
Senior Research Scientist, Supervisor, Group Leader of Sensory Resonance Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories.

He received a B.E. and M.E. in environmental and sanitary engineering from Kyoto University in 1991 and 1993, and a Ph.D. in auditory perception from University of Cambridge, UK, in 1996. He conducted postdoctoral studies in the USA between 1996 and 2001. As a postdoctoral associate at Kresge Hearing Research Institute at the University of Michigan, USA, he conducted electrophysiological studies on sound localization, specifically the representation of auditory space in the auditory cortex. He joined NTT Communication Science Laboratories in 2001. Since then, he has been involved in studies on auditory-space representation in the brainstem, assessing basic hearing functions, and the salience of auditory objects or events. In addition, as the group leader of the Sensory Resonance Research Group, he is managing various projects exploring mechanisms that underlie explicit and implicit communication between individuals. He was the principal investigator of a MIC SCOPE commissioned research project on auditory salience. He is a member of the Acoustical Society of America, the Acoustical Society of Japan (ASJ) (member of the Executive Council), the Association for Research in Otolaryngology (ARO), and the Japan Neuroscience Society (JNS).

Shimpei Yamagishi
Ph.D. student, Department of Information Processing, Tokyo Institute of Technology (External collaborative program with NTT).

He received a B.E. in astrophysics from Tokyo University of Science in 2011 and an M.E. in auditory neuroscience from Tokyo Institute of Technology in 2013. He has been studying neural processing for auditory object formation at the subcortical level. He received the 2012 Student Best Presentation Award from ASJ. He was granted a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science (2015–). He is a member of ASJ, JNS, and ARO.

Hsin-I Liao
Research Associate, Sensory Resonance Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories.

She received a B.S. and Ph.D. in psychology from National Taiwan University in 2002 and 2009. She joined NTT Communication Science Laboratories in 2012 and has been studying auditory salience, music preference, and preference of visual images. She has also explored the use of pupillary response recording to correlate human cognitive functions such as auditory salience and preference decision. During 2007–2008, she was a visiting student at California Institute of Technology, USA, where she studied visual preference using recorded eye movements and visual awareness using transcranial magnetic stimulation. She received a Best Student Poster Prize at the Asia-Pacific Conference on Vision (APCV) in 2008, a Travel Award by the Association for the Scientific Study of Consciousness (ASSC) in 2011, and a Registration Fee Exemption Award by the International Multisensory Research Forum (IMRF) in 2011. She is a member of the Vision Sciences Society (VSS), ARO, and JNS.

Makoto Yoneya
Researcher, Sensory Resonance Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories.

He received a B.E. and M.Sc. in engineering from the University of Tokyo in 2010 and 2012. He joined NTT Communication Science Laboratories in 2012 and has been studying biological signal processing, especially of eye movements. He is also interested in decoding people's thoughts based on brain or neural activity using machine learning methods and has researched decoding of the 'internal voice' using magnetoencephalography signals and multi-class SVM (support vector machine). He is also studying auditory signal processing and is currently developing a mathematical model of auditory salience. He received the 2011 Best Presentation Award from the Vision Society of Japan. He is a member of ASJ, JNS, and ARO.

Sho Otsuka
Research Associate, Sensory Resonance Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories.

He received an M.A. and Ph.D. in environmental science from the University of Tokyo in 2011 and 2014. He joined NTT Communication Science Laboratories in 2014 and has been studying the assessment of hearing difficulties using otoacoustic emissions (OAEs) and frequency-following response. He is also investigating hidden hearing loss, which is a form of hearing disorder that does not show up in conventional auditory tests, by evaluating inner-ear mechanical properties using OAEs. He received a Student Presentation Award and an Awaya Prize Young Researcher Award from ASJ in 2012 and 2013, respectively. He is a member of ASJ and ARO.

Makio Kashino
Senior Distinguished Scientist/Executive Manager of Human Information Science Laboratory, NTT Communication Science Laboratories.

He received a B.A., M.A., and Ph.D. in psychophysics from the University of Tokyo in 1987, 1989, and 2000. He joined NTT in 1989. From 1992 to 1993, he was a Visiting Scientist at the University of Wisconsin (Prof. Richard Warren's laboratory), USA. Currently, he is a Visiting Professor in the Department of Information Processing, Tokyo Institute of Technology (2006–), and PI (principal investigator) of a JST CREST project on implicit interpersonal information (2009–). He has been investigating functional and neural mechanisms of human cognition, especially auditory perception, cross-modal and sensorimotor interaction, and interpersonal communication through the use of psychophysical experiments, neuroimaging, physiological recordings, and computational modeling.

Deep Learning Based Distant-talking Speech Processing in Real-world Sound Environments

Shoko Araki, Masakiyo Fujimoto, Takuya Yoshioka, Marc Delcroix, Miquel Espi, and Tomohiro Nakatani

Abstract

This article introduces advances in speech recognition and speech enhancement techniques with deep learning. Voice interfaces have recently become widespread. However, their performance degrades when they are used in real-world sound environments, for example, in noisy environments or when the speaker is some distance from the microphone. To achieve robust speech recognition in such situations, we must make progress in further developing various speech processing techniques. Deep learning based speech processing techniques are promising for expanding the usability of a voice interface in real and noisy daily environments.

Keywords: deep learning, automatic speech recognition, speech enhancement

1. Introduction

In recent years, the use of voice-operable smartphones and tablets has become widespread, and their usefulness has been widely recognized. When a user speaks carefully into a terminal, that is, a microphone(s) (**Fig. 1(a)**), his/her voice is usually accurately recognized, and the device works as expected.

On the other hand, there is a growing need for voice interfaces that can work when a user speaks at a certain distance from the microphones. For example, when we record the discussion in a meeting, as shown in **Fig. 1(b)**, we may want to employ a terminal on the table and avoid the use of headset microphones. Furthermore, when users talk to voice-operated robots or digital signage, the users would talk to them from a certain distance.

However, the current speech recognition accuracy of voice-operable devices is generally insufficient when the speaker is far away from the microphone. This is because of the considerable effect of noise and reverberation and because the users speak freely with little awareness of the microphones when the microphones are some distance away. We are therefore studying distant speech recognition and working on speech enhancement and speech recognition techniques to expand the usability of a voice interface in real-world sound environments.

There are two main factors that degrade the automatic speech recognition of distant speech; (1) the quality of speech recorded with a distant microphone is severely degraded by background noise, for example, air conditioners and room reverberation. Moreover, in a multi-person conversation, the speakers' voices sometimes overlap. (2) As the users speak freely without regard to the microphones, their utterances become fully spontaneous and therefore tend to include ambiguous pronunciations and abbreviations. Speech enhancement techniques are essential in order to cope with such complex difficulties, and these include noise reduction, reverberation reduction (dereverberation), speech separation, and spontaneous speech recognition techniques.

Close microphone(s)

(a) Current status

Distant microphone(s)

(b) Our goal

Fig. 1. Current and future status of voice interfaces.

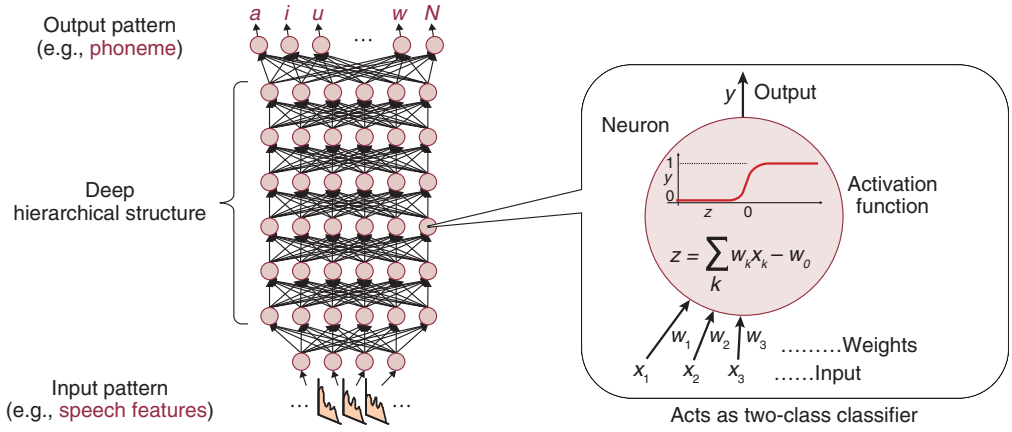


Fig. 2. Acoustic model with deep neural network (DNN).

2. Deep learning in speech processing

We have been studying the aforementioned speech processing techniques in order to achieve distant speech recognition in the real world. In recent years, we have been working on speech processing methods based especially on deep learning. Deep learning is a machine learning method that uses a deep neural network (DNN), as shown in **Fig. 2**. Deep learning has recently come under the spotlight because in 2011 and 2012 it was shown to outperform conventional techniques in many research fields including image recognition and compound activity prediction. High performance has also been achieved with deep learning in speech recognition tasks, and therefore, deep learning based speech processing techniques have been intensively researched in recent years.

In 2011, we began working on deep learning based techniques for automatic recognition of spontaneous speech [1]. It should be noted that a deep learning based real-time speech recognizer developed by NTT Media Intelligence Laboratories has already been released [2]. We have also proven that deep learning improves speech enhancement techniques such as noise reduction when deep learning is effectively leveraged. The remainder of this article describes our speech recognition and speech enhancement techniques that employ deep learning.

3. Speech recognition with deep learning

General automatic speech recognition techniques translate features into phonemes, phonemes into words, and words into sentences, by respectively using an acoustic model, a pronunciation dictionary, and a language model (**Fig. 3**). Originally, deep

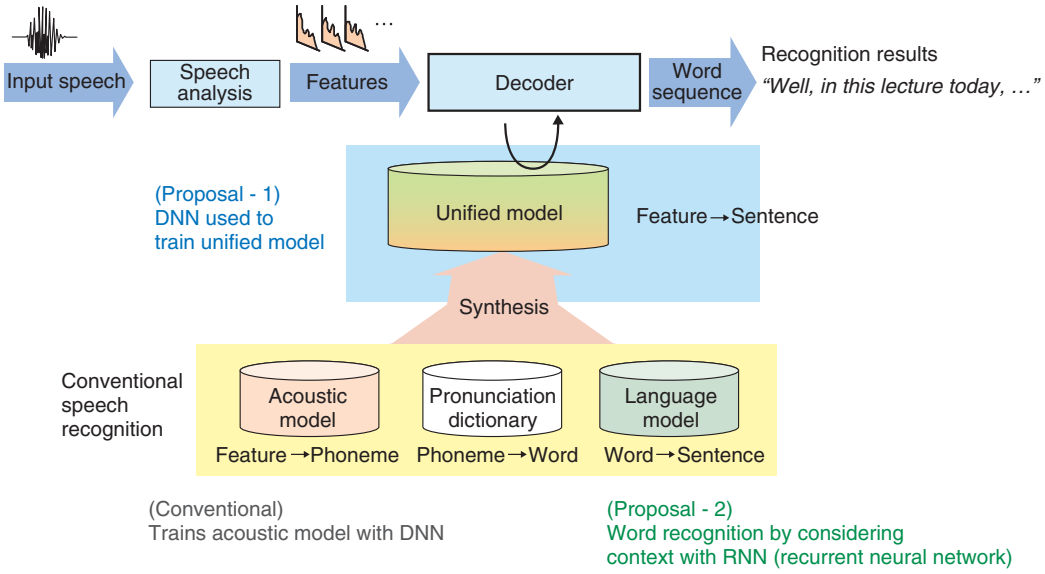


Fig. 3. Speech recognition process.

learning based speech recognition employed a DNN to achieve accurate acoustic modeling, and it outperformed conventional speech recognition techniques that do not use deep learning.

The aforementioned acoustic model, pronunciation dictionary, and language model are usually trained separately, so it has been difficult to consider the interaction between the phonetic and linguistic factors that are present in spontaneous speech. To address these complex factors, we proposed synthesizing the three models into a unified model (Fig. 3) and optimizing it by using a DNN [1]. We demonstrated that this unified model achieves highly accurate spontaneous speech recognition [1].

Moreover, we showed that a recurrent neural network (RNN), which is also a deep learning technique, in a language model provides further improvement in performance. An RNN based language model is effective for spontaneous speech recognition because its ability to hold the history of words enables us to recognize speech by considering a longer context. However, it is generally difficult to achieve fast automatic speech recognition while maintaining the complete contextual history. We therefore proposed an efficient computational algorithm for maintaining contexts and achieved fast and highly accurate automatic speech recognition [3].

The word error rates (WERs) in English lecture speech recognition are shown in **Fig. 4**. *Without DNN* indicates the WER before deep learning was employed. The *DNN acoustic model* shows the large effect of deep learning. We can also see that the *unified DNN*, where the unified model is optimized with a DNN, outperforms the conventional *DNN acoustic model*. Moreover, the *RNN language model* achieves the best performance, which is more than 4 points better than the conventional *DNN acoustic model*. The appropriate use of deep learning techniques significantly improves spontaneous speech recognition performance.

4. Speech enhancement with deep learning

Deep learning also helps to improve speech enhancement performance. This section introduces two noise reduction techniques: a method for use with multiple microphones and a method for use with a single microphone.

The first approach estimates the features of noise-reduced speech by using a DNN (**Fig. 5(a)**). Pairs consisting of clean and noisy speech signals are used to train the DNN to translate noisy speech features into clean speech features. The trained DNN is then used to estimate noise-reduced features when the input consists of noisy features. This method was originally used for noise reduction with a single microphone, however, its extension to multi-microphone use was not obvious. We found that we can improve noise reduction performance by inputting additional features estimated with multi-microphone

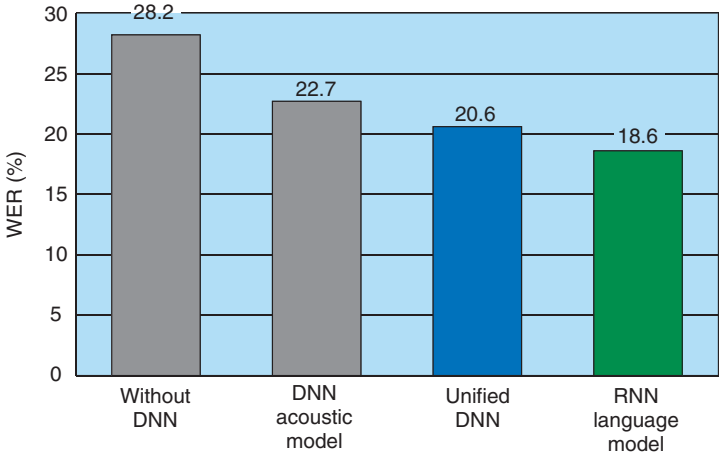


Fig. 4. Word error rates in English lecture speech recognition.

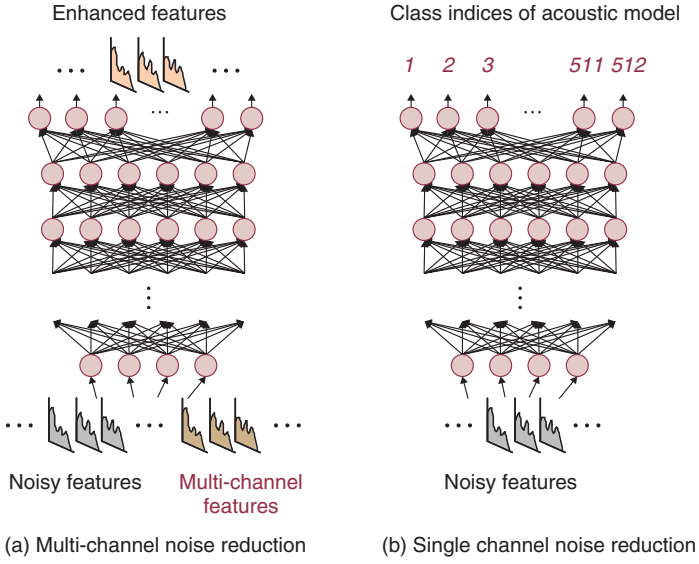


Fig. 5. Examples of DNN-based noise reduction.

observations into a DNN. We also found that the probability of speech existing at each time-frequency slot, which can be estimated with a microphone array technique, provides us with an effective additional feature [4]. The results of an evaluation conducted under living room noise conditions (PASCAL CHiME challenge task) revealed the superiority of our proposed approach. Specifically, we obtained a reduced WER of 8.8% by using the proposed multi-microphone features compared to a value of 10.7% without them.

The second noise reduction approach is for cases where we can use just a single microphone. This method is applied to calculate noise reduction filter coefficients by using probabilistic models of clean speech and noise without speech. Here, accurate model estimation is important for accurate filter design. We showed that DNN-based clean speech model estimation (**Fig. 5(b)**) achieves high-performance noise reduction [5]. Specifically, we constructed a clean speech model with a set of probabilistic models and utilized a DNN to discriminate the

most suitable model for generating the observed noisy speech. With this proposed noise reduction approach, we obtained an improved WER of 19.6% for a noisy speech database (AURORA4), whereas the WER was 23.0% with a conventional method without a DNN.

It is worth mentioning that we do not use a DNN for noise model estimation. This is because it is difficult to obtain a sufficient quantity of noise data for DNN training due to the wide range of variations and momentary fluctuations of noise in the real world. With the proposed method, we estimate the noise model using an unsupervised method, and we simultaneously use a DNN for clean model selection. This approach achieves high-performance noise reduction in real-world sound environments by flexibly considering the variation of noisy signals.

5. Outlook

We believe that distant-talking speech processing is a key technology for expanding the usability of voice interfaces in actual daily life. In particular, conversational speech recognition and communication scene analysis in real-world sound environments are techniques that meet the needs of the times. These techniques should make a significant contribution to artificial intelligence (AI) speech input, which has recently attracted renewed interest for applications such as minute-taking systems in business meetings, intelligent home electronics, and a human-robot dialogue system for use in shopping centers. For these purposes, we need a highly accurate distant speech recognition technique that works in noisy environments. In addition, techniques for identifying the current speakers and for understanding what is going on around the AI device by recognizing, for example, environmental sound events [6], are also becoming more important. We are continuing to work on the development of essential techniques for distant-talking speech processing in order to expand the capabilities of voice interfaces to their fullest extent.

References

[1] 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

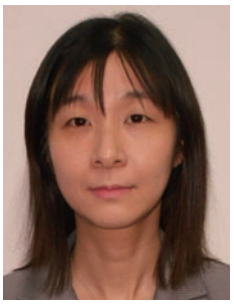
[2] NTT IT CORPORATION News Release, Nov. 11, 2014 (in Japanese). http://www.ntt-it.co.jp/press/2014/1111/

[3] T. Hori, Y. Kubo, and A. Nakamura, “Real-time One-pass Decoding with Recurrent Neural Network Language Model for Speech Recognition,” Proc. of ICASSP 2014 (2014 IEEE International Conference on Acoustics, Speech, and Signal Processing), pp. 6364–6368, Florence, Italy, May 2014.

[4] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, “Exploring Multi-channel Features for Denoising-autoencoder-based Speech Enhancement,” Proc. of ICASSP 2015, pp. 116–120, Brisbane, Australia, Apr. 2015.

[5] M. Fujimoto and T. Nakatani, “Feature Enhancement Based on Generative-discriminative Hybrid Approach with GMMs and DNNs for Noise Robust Speech Recognition,” Proc. of ICASSP 2015, pp. 5019–5023, Brisbane, Australia, Apr. 2015.

[6] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, “Feature Extraction Strategies in Deep Learning Based Acoustic Event Detection,” Proc. of Interspeech 2015, pp. 2922–2926, Dresden, Germany, Sept. 2015.



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

She received a B.E. and M.E. from the University of Tokyo in 1998 and 2000, and a Ph.D. from Hokkaido University in 2007. Since joining NTT in 2000, she has been conducting research on acoustic signal processing, array signal processing, blind source separation, meeting diarization, and auditory scene analysis.

She was a member of the organizing committee of ICA 2003, IWAENC 2003, WASPAA 2007, and the evaluation co-chair of SiSEC 2008, 2010, and 2011. Since 2014, she has been a member of the Institute of Electrical and Electronics Engineers (IEEE) Signal Processing Society Audio and Acoustics Technical Committee. She received the 19th Awaya Prize from the Acoustical Society of Japan (ASJ) in 2001, the IWAENC Best Paper Award in 2003, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004 and 2014, the Academic Encouraging Prize from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2006, the Itakura Prize Innovative Young Researcher Award from ASJ in 2008, and the Young Scientists' Prize of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2014. She is a member of IEEE, IEICE, and ASJ.

Masakiyo Fujimoto
Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

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

Takuya Yoshioka
Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received his B.Eng., M.Inf., and Ph.D. in informatics from Kyoto University in 2004, 2006, and 2010. In 2005, he interned at NTT, where he conducted research on dereverberation. Since joining NTT in 2006, he has been working on the development of algorithms for noise robust speech recognition, speech enhancement, and microphone arrays. During 2013–2014, he was a Visiting Scholar at the University of Cambridge, Cambridge, UK. He has been a part-time lecturer at Doshisha University, Kyoto, since 2015. He received the Awaya Prize Young Researcher Award and the Itakura Prize Innovative Young Researcher Award from ASJ in 2010 and 2011, respectively, and the Young Researcher's Award in Speech Field from IEICE ISS in 2011.

Marc Delcroix
Senior Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received an M.E. from the Free University of Brussels, Belgium, and the Ecole Centrale Paris, France, in 2003 and a Ph.D. from the Graduate School of Information Science and Technology, Hokkaido University, Japan, in 2007. He joined NTT Communication Science Laboratories in 2010. He is also a Visiting Lecturer in the Faculty of Science and Engineering of Waseda University, Tokyo. From 2004 to 2008, he was a research associate at NTT Communication Science Laboratories. From 2008 to 2010, he worked at Pixela Corporation developing software for digital television. His research interests include robust speech recognition, speech enhancement, and speech dereverberation. He was one of the organizers of the REVERB challenge 2014. He received the 2005 Young Researcher Award from the Kansai section of ASJ, the 2006 Student Paper Award from the IEEE Kansai section, and the 2006 Sato Paper Award from ASJ. He is a member of IEEE and ASJ.

Miquel Espi
Research Associate, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E. in computer science from Universidad Politecnica de Valencia, Spain, in 2006, an M.E. in information science from Kagoshima University in 2010, and a Ph.D. in information science and technology from the University of Tokyo in 2013. He joined NTT Communication Science Laboratories in 2013 and has been researching characterization and classification of acoustic events in the context of conversation scene analysis. His current research interests include acoustic scene analysis, acoustic signal processing, and social dynamics. He is a member of IEEE, IEEE Signal Processing Society, and ASJ.

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

He received his B.E., M.E., and Ph.D. from Kyoto University in 1989, 1991, and 2002. He joined NTT Basic Research Laboratories in 1991 and moved to NTT Communication Science Laboratories in 2001. During 2005–2006, he was a Visiting Scholar at Georgia Institute of Technology, USA. Since 2008, he has been a Visiting Assistant Professor in the Department of Media Science, Nagoya University, Aichi. His research interests include speech enhancement technologies for intelligent human-machine interfaces. He received the 1997 JSAI (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. During 2009–2014, he was a member of the IEEE Signal Processing Society Audio and Acoustics Technical Committee (AASP-TC), and he has been an associate member of the AASP-TC since 2015. He has also served as Chair of the review subcommittee of AASP-TC, Associate Editor of IEEE Transactions on Audio, Speech, and Language Processing, Chair of the IEEE Kansai Section Technical Program Committee, Technical Program co-Chair of IEEE WASPAA-2007, and as a member of the IEEE Circuits and Systems Society Blind Signal Processing Technical Committee. He is a member of IEEE, IEICE, and ASJ.

Yu bi Yomu: A New Text Display System Using Tracing Behavior

Kazushi Maruya and Junji Watanabe

Abstract

NTT Communication Science Laboratories is researching a text display system called Yu bi Yomu, in which the appearance of text changes dynamically in response to the user's finger-tracing behavior. Research on digital text display has so far centered on discussions on how to achieve the feeling of using paper. However, digital text display has the potential to surpass attempts at simply imitating the paper medium by exploiting digital features in order to bring about major changes in the way that reading itself is performed. This article provides an overview of the Yu bi Yomu system and introduces the advantages of using this method.

Keywords: text display technology, interactive interface, finger-tracing reading

1. Introduction

Paper has been the main medium for presenting text for a very long time. However, recent progress in digital media technology is rapidly increasing the opportunities for displaying text on computer displays (digital text displays) instead of printing text on paper. Digital text display can do more than just simplify the handling of huge amounts of textual information. It also has the potential to bring about major changes in the way that reading itself is performed, that is, in reading behavior, by exploiting the features of digital devices.

In principal, the content of text presented on paper does not change. A digital device, however, enables the dynamic presentation of information [1–4]. In this regard, the market for devices equipped with a touch-response function such as tablet computers and smartphones has been growing rapidly in recent years. A key feature of these digital devices is the adoption of an interface by which the user can directly touch and manipulate what is being displayed. Compared to keyboard or mouse operations, using one's fingers or a touch pen is highly intuitive and interactive. Such methods of direct manipulation have also come to be used for perusing text documents, but most of those methods have been focused on reproducing the sense of manipulating media having fixed text, and the paper medium in particular. The benefits of performing direct operations on characters and text as symbols that convey information have hardly been considered to date.

NTT Communication Science Laboratories is researching a text display system called *Yu bi Yomu* [5, 6] that makes use of the dynamic-display and touch-based direct-manipulation features of tablet devices (**Fig. 1**). This system displays text very faintly when the reader is not interacting with the display. However, if the reader begins to trace the characters of that text with his or her finger, the system will gradually increase the contrast of those characters, which is known as fade-in. At this time, the high-contrast characters can be left in that state, but it is also possible to configure the settings so the characters gradually fade out after a specific length of time. In either case, the reader proceeds to read the text while the characters of that text become increasingly visible by finger tracing. In addition, dynamic text display enables intuitive manipulation of such features of reading text as emotional response and intonation that are difficult for static text display to explicitly represent.

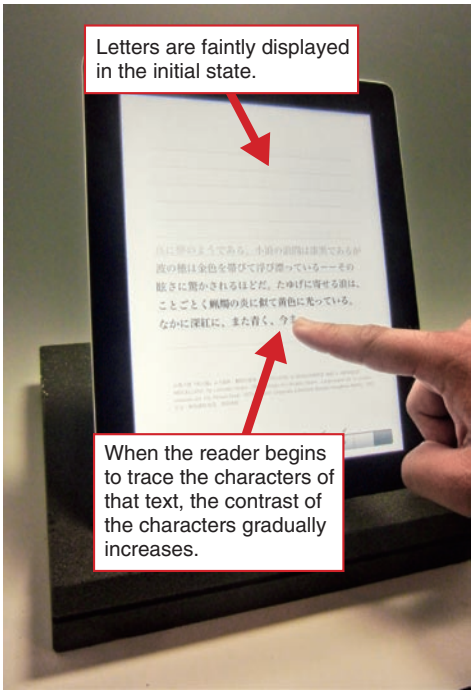


Fig. 1. Yu bi Yomu system.

2. Basic structure of software using the Yu bi Yomu system

We are currently implementing the Yu bi Yomu system as software running on commercially available tablet computers. We are also developing software for a variety of devices running under iOS[*1], Android[*2] OS (operating system), Mac OS[*3], and Windows[*4] personal computers (touch-panel displays required), although functions will differ somewhat among these devices. This software assumes the preparation of text files that use a special format to manipulate the text that the user would like to display with the Yu bi Yomu system (**Fig. 2(a)**). This format is a type of Extensible Markup Language (XML)[*5] and describes the text to be displayed and symbols called *tags* in much the same way that web pages are prepared. Preparing text files in this format (Yu bi Yomu files) enables the user to change the text to be displayed as well as the temporal profile of elements such as the text font and font size without having to modify the software program. For example, specifying the following four parameters controls the temporal change in the appearance of text when finger tracing (Fig. 2(a)).

(1) Time from user touching the screen to start of text fade-in
(2) Time from start of text fade-in to maximum contrast
(3) Time that text is held at maximum contrast
(4) Time from start of text fade-out back to original state

We found that manipulating these parameters describing the temporal profile of text display can change the reader's overall impression of reading a document, as indicated by the words *warm* and *soft*, and *cool* and *stiff* in **Fig. 2(b)** [6].

The Yu bi Yomu system can also record the user's tracing behavior. The user can start and stop this recording by pressing a software button and can make use of the recorded tracing behavior without having to do any programming or editing. In particular, the data obtained by recording tracing behavior can be used to prepare animation in which text automatically rises up based on a time series of finger-tracing actions. Additionally, recorded data may be saved in a tab-separated text file and exported to spreadsheet software such as Excel[*6] to enable reading behavior to be analyzed for educational purposes.

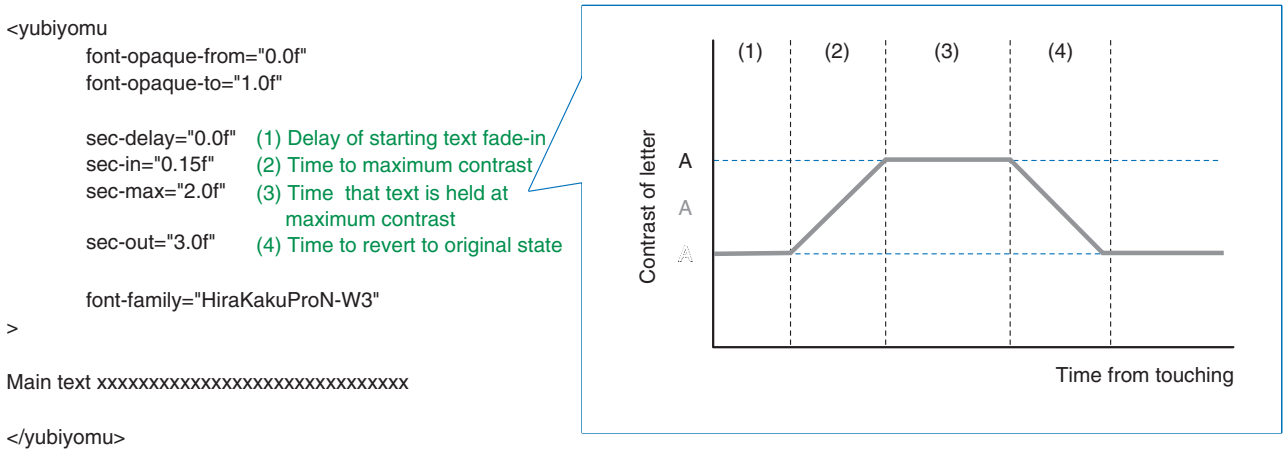
3. Usage scenarios for the Yu bi Yomu system

The Yu bi Yomu system can be applied to the text-communication and education fields (**Fig. 3**).

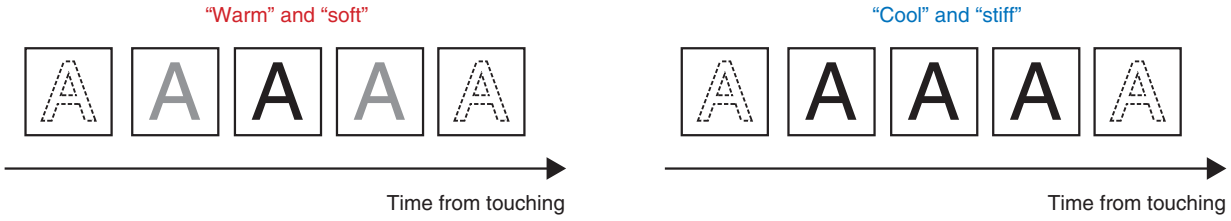
3.1 Application to text communication

We consider that incorporating software using the Yu bi Yomu system in text-communication tools such as email and messaging apps can facilitate smooth communication. For example, text-based animation using tracing data can enable a user to add nuance and subtle emotions to short messages. Animated text using tracing data can also be combined with a function for recording audio during finger tracing so that an animated message can be created that combines speech and moving text.

*1 iOS is a registered trademark of Cisco Systems Inc. in the United States and other countries and is used by Apple Inc. under license.
*2 Android is a registered trademark of Google Inc. in the United States and other countries.
*3 Mac OS is a trademark of Apple Inc. in the United States and other countries.
*4 Windows is a registered trademark of Microsoft Corporation in the United States and other countries.
*5 XML: The name of a language or its specifications for describing instructions related to the logical structure or form of text together with the body of that text all in a text file.
*6 Excel is a registered trademark of Microsoft Corporation in the United States and other countries.



(a) Examples of parameters that control the temporal change in appearance of text



(b) Impression changes in reading the text by manipulating temporal parameters

Fig. 2. Basic structure of software using the Yu bi Yomu system.

Animation in which text automatically flows on a display has come to be used in many everyday scenarios such as television captions, karaoke systems, and electronic billboards. Experts have traditionally prepared such animation using computers and specialized software, but the Yu bi Yomu system makes it easy for even general users to create animated text for smartphones and other devices.

Furthermore, in most animated text prepared by computer, the speed at which text appears is fixed. In contrast, animated text based on human tracing includes text whose appearance accelerates and decelerates in an uneven manner. Such complex displays of text can express intonation in a message and convey the sender's individuality and presence [7, 8], which cannot easily be achieved with traditional computer systems.

3.2 Application to education

We consider that the Yu bi Yomu system could also be applied with good effect in the field of education. Compared to silent reading, reading while tracing accompanies a physical action (finger tracing) that can be observed by other people. Moreover, as the position of words being read basically agrees with the position of the finger while tracing, examining a record of finger tracing makes it easy to learn how that person is progressing in reading that text. In this way, the Yu bi Yomu system makes it relatively simple to visualize the way in which someone reads text.

For example, in a classroom setting, using the Yu bi Yomu system to prepare teaching material makes it easy to visualize interactive behavior while reading, which has the potential to improve the quality of instruction. We conducted an experiment in collaboration with NTT-ME in the use of Yu bi Yomu teaching material. In this experiment, we introduced teaching material using the Yu bi Yomu system in classroom-based training targeting ordinary employees, and we compared the results with instruction using the same material but displayed statically in the usual way on a tablet device [9]. We found that the average

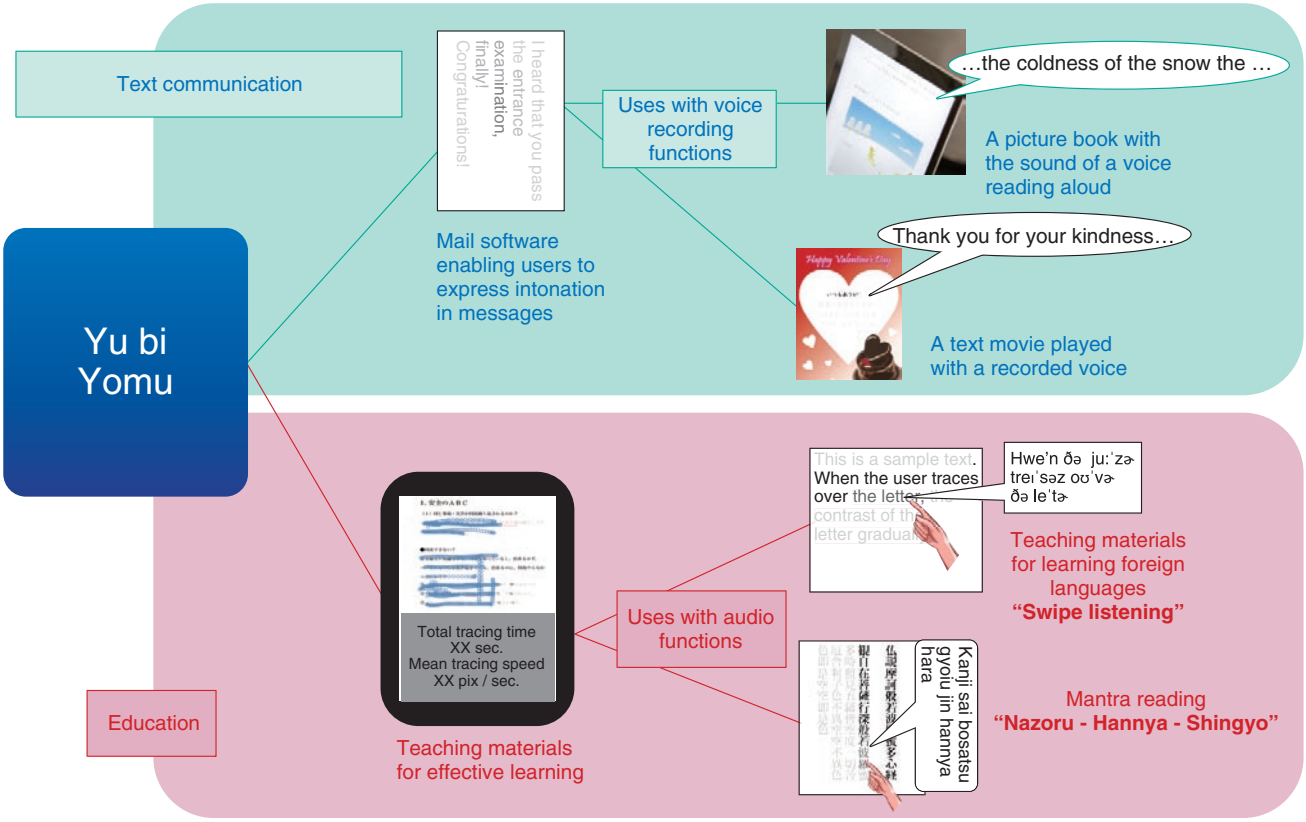


Fig. 3. Application of the Yu bi Yomu system.

score in a memory test taken after training was about 10% higher in the group receiving instruction using Yu bi Yomu material than in the group using material presented in the conventional way. There are a number of possibilities as to why test results improved in this way, and further research is needed to arrive at a detailed explanation. At the least, however, the results of this experiment have shown that the quality of instruction can be improved by using the Yu bi Yomu system for certain types of teaching material and classroom settings.

Furthermore, on converting the tracing data obtained from learners who used Yu bi Yomu teaching materials into graphic images and examining the results, it was found that the way in which reading was approached could differ greatly between people. Examples of finger tracing between two learners—one who did well in the recollection test and one who did poorly—are shown in **Fig. 4**. It is obvious which of the two approached the material in a more serious manner.

Text display using the Yu bi Yomu system can also be combined with a text-to-speech function to prepare teaching material on foreign languages and ancient texts such as sutras that may be difficult to read on one's own. For example, software to read text out loud while finger tracing could be created by preparing a text-to-speech file beforehand and changing the generated speech in unison with the text being traced. Likewise, educational software with a text-to-speech function could provide a sense of self-improvement as one learns how to read text that one could not otherwise read on one's own. This approach holds the possibility of keeping the learner motivated and focused with a desire to continue learning.

4. Future outlook

Incorporation of the Yu bi Yomu system introduced in this article in the software used in popular digital devices such as smartphones and tablets is expected to expand the communication functions provided by those devices. Furthermore, combining the software with a text-to-speech function should make it

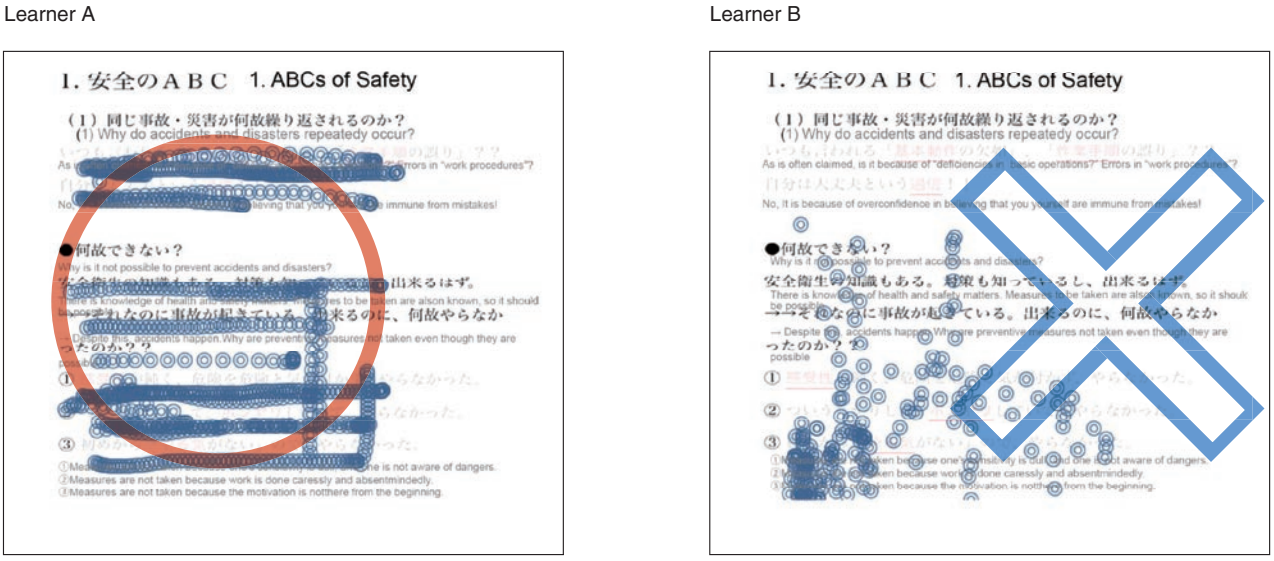


Fig. 4. Examples of finger tracing.

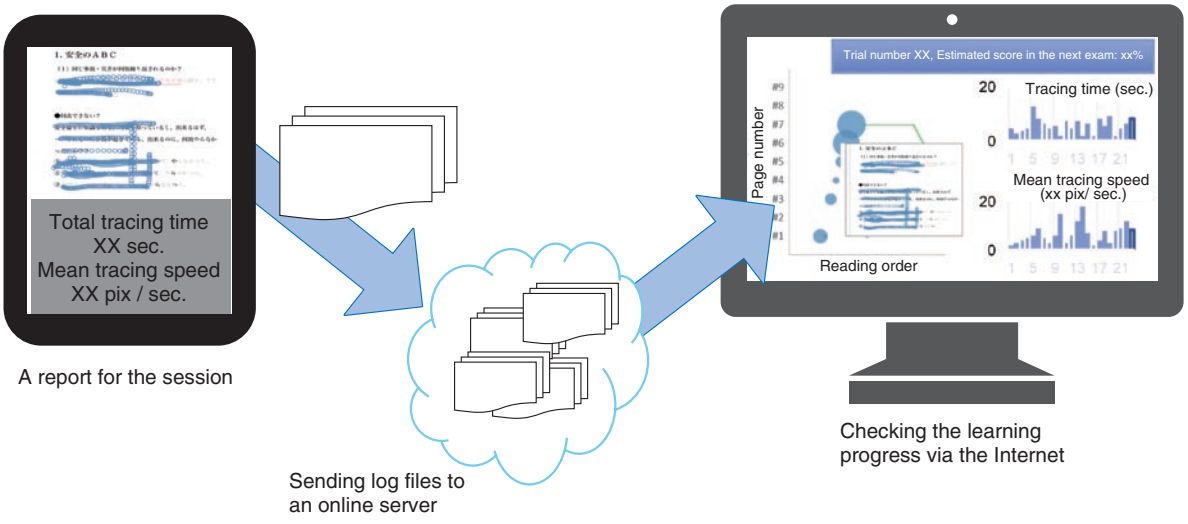


Fig. 5. Application to large-scale learning systems.

applicable to picture books read out loud by a familiar voice. In addition, collecting a large amount of tracing data over time may make it possible to uncover key relationships between tracing and learning conditions. If so, we can expect the Yu bi Yomu system to be useful in large-scale learning systems (**Fig. 5**).

The Yu bi Yomu system focuses on the temporal and interactive characteristics associated with the digital display of information. It is an attempt at incorporating the temporal expression of emotional response and intonation common to the spoken language in textual expression using tracing behavior. In this research, it is not our aim to reproduce on digital devices textual expression as achieved on the medium of paper. Rather, we seek to exploit the special features and formats of digital devices to open up new possibilities in communication.

References

[1] J. C. Lee, J. Forlizzi, and S. E. Hudson, "The Kinetic Typography Engine: an Extensible System for Animating Expressive Text," Proc. of UIST 2002 (15th annual ACM Symposium on User Interface Software and Technology), ACM Press (2002), pp. 81–90, Paris, France, Oct. 2002.

[2] J. Forlizzi, J. C. Lee, and S. E. Hudson, "The Kinedit System: Affective Messages Using Dynamic Texts," Proc. of CHI (Conference on Human Factors in Computing Systems) 2003, ACM Press (2003), pp. 377–384, Ft. Lauderdale, Florida, USA, Apr. 2003.

[3] Y. Y. Wong, "Temporal Typography: a Proposal to Enrich Written Expression," Proc. of CHI 1996, ACM Press (1996), pp. 408–409, Vancouver, Canada, Apr. 1996.

[4] G. Möhler, M. Osen, and H. Harrikari, "A User Interface Framework for Kinetic Typography-enabled Messaging Applications," Proc. of CHI 2004, ACM Press (2004), pp. 1505–1508, Vienna, Austria, Apr. 2004.

[5] K. Maruya, M. Uetsuki, H. Ando, and J. Watanabe, ""Yu bi Yomu": Interactive Reading of Dynamic Text," Proc. of MM 2012 (20th ACM International Conference on Multimedia), ACM Press (2012), pp. 1499–1500, Nara, Japan, Apr. 2012.

[6] K. Maruya, M. Uetsuki, H. Ando, and J. Watanabe, "Dynamic Text Display Using Finger Trailing," IPSJ Journal, Vol. 54, No. 4, pp. 1507–1517, 2012 (in Japanese).

[7] F. Heider and M. Simmel, "An Experimental Study of Apparent Behavior," The American Journal of Psychology, Vol. 57, No. 2, pp. 243–259, 1944.

[8] P. D. Tremoulet and J. Feldman, "The Influence of Spatial Context and the Role of Intentionality in the Interpretation of Animacy from Motion," Perception and Psychophysics, Vol. 68, No. 6, pp. 1047–1058, 2006.

[9] K. Maruya, J. Watanabe, H. Takahashi, and S. Hashiba, "A Learning System Utilizing Learners' Active Tracing Behaviors," Proc. of LAK15 (5th International Conference on Learning Analytics and Knowledge), pp. 418–419, New York, USA, Mar. 2015.

Kazushi Maruya

Senior Research Scientist, Sensory Representation Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories*.

He received an M.A. and a Ph.D. in psychology from the University of Tokyo in 2001 and 2005. He joined NTT Communication Science Laboratories in 2008. He was a researcher in the Intelligent Modeling Laboratory, the University of Tokyo, from 2004 to 2005, and a visiting scientist in the Department of Psychology, Vanderbilt University, Nashville, USA, from 2006 to 2008. His current research interests include visual motion processing for perception, especially for natural scene perception, and human computer interface design for reading digital text.

* He moved to Research Planning Department of NTT Science and Core Technology Laboratory Group on August 1, 2015.

Junji Watanabe

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

He received a Ph.D. in information science and technology from the University of Tokyo in 2005. His academic work has been published in scientific journals in the field of neuroscience and interface technologies. He has also presented his work at technology showcases, science museums, and art festivals such as at SIGGRAPH (2006–2009, 2014) and Ars Electronica (2002, 2004, 2007–2015). His research is focused on cognitive science and communication devices with applied perception.

Combinatorial Optimization Using Binary Decision Diagrams

Masaaki Nishino, Norihito Yasuda, Tsutomu Hirao, Shin-ichi Minato, and Masaaki Nagata

Abstract

Combinatorial optimization is being used to solve a wide range of real world tasks, but its application requires that we formulate the task as an optimization problem for which efficient methods for solving the problem exist. However, sometimes task-specific constraints prevent us from formulating the task as an easy-to-solve optimization problem. In this article, we present a new algorithm for solving combinatorial optimization problems by using a binary decision diagram (BDD), a data structure for representing a Boolean function as a compact graph. Our method can efficiently solve constraint-added variants of a class of optimization problems by representing the constraints with a BDD or zero-suppressed BDD (ZDD) and then applying an efficient dynamic programming algorithm.

Keywords: combinatorial optimization, binary decision diagram, natural language processing

1. Introduction

In daily life, we make many decisions ranging from important decisions on business matters to choosing what to eat for lunch. How do people choose one action from all possible actions? It seems natural to assume that people rely on some criteria for selecting their action, rather than selecting it at random. Here, we set the assumption that every possible action can be scored, a value that represents how *good* that action is. Under this assumption, making a decision can be regarded as solving the problem of finding an action with a maximum score. This is called an *optimization problem*. If all possible candidates are represented as the assignment of discrete values on variables, then the problem is called a *combinatorial optimization problem*. In addition to decision making in daily life, techniques that involve solving combinatorial optimization problems are used in performing various computer science tasks.

2. Solving real world tasks with combinatorial optimization techniques

To solve a real world task using combinatorial optimization, we first have to formulate the task as an optimization problem, and then solve the problem. The formulation consists of (i) setting *constraints* that all possible actions must satisfy and (ii) designing *an objective function* that takes a possible choice and returns the score that represents how good the choice is. Let the choices satisfying the constraints be *possible solutions*, and let the possible solution that maximizes the objective function be *the optimal solution*.

We can solve several tasks by formulating them as combinatorial optimization problems. As a very simple example, let us consider the situation in which someone wants to buy some food at a grocery store under the constraint that the total price spent must be within 300 yen. We assume that the store sells n kinds of items, and every item has a satisfaction score. The score represents the degree to which the user will be satisfied upon buying the item(s). Under this assumption, the task of buying food is formulated as the combinatorial optimization problem of finding the best combination that maximizes the sum of satisfaction scores while keeping the total price within 300 yen (**Fig. 1**). We can get the best combination of items within 300 yen by solving the optimization problem.

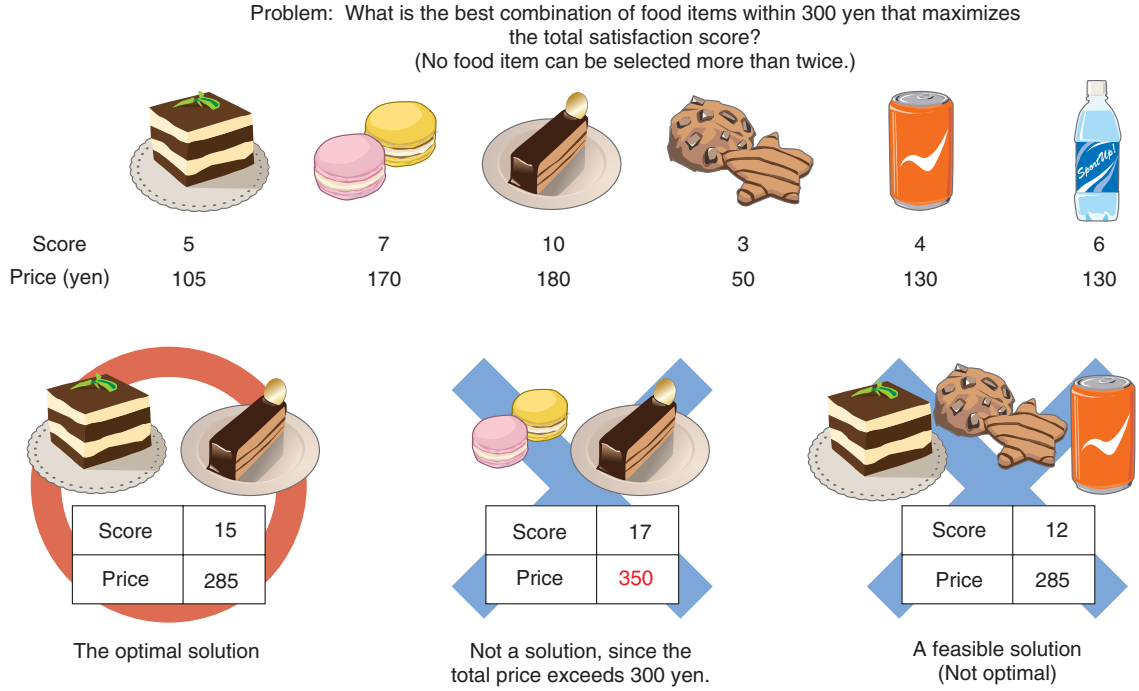
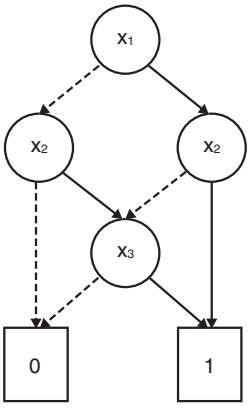


Fig. 1. A combinatorial optimization problem.

The degree of difficulty in solving an optimization problem depends on the problem. Thus, it is important to formulate a task as a combinatorial optimization problem for which efficient solution algorithms exist. The knapsack problem and the shortest path problem are typical examples of combinatorial optimization problems, and their algorithms are often applied. Unfortunately, such algorithms fail to offer full problem coverage due to certain limitations. For example, the above example of buying items can be formulated as a knapsack problem, and we can find the optimal solution in linear time ($n \times L$, where n is the number of items and L is the budget). However, the knapsack algorithm fails to consider relationships between selected items. This means that conditions such as *should not buy item A and item B at the same time*, or *should buy at least item C or D* cannot be considered when finding the optimal solution.

Many task-specific conditions must be considered when solving real world tasks, and these constraints prevent the tasks from being formulated as easy-to-solve combinatorial optimization problems. If no efficient algorithm exists for solving a problem, we have two choices; the first is to develop a new solution algorithm, and the second is to use a general-purpose optimization software package. Taking the former approach is unrealistic because it is virtually impossible for non-experts to design an efficient algorithm. Taking the latter approach is relatively easy, and off-the-shelf software can be used to solve various kinds of optimization problems. However, this approach has a shortcoming in that we cannot estimate the time required for solving the problem beforehand.

We have developed a new efficient optimization algorithm for solving a class of optimization problems. This class consists of problems that can be formulated as the addition of constraints that consider discrete relationships among variables to easy-to-solve optimization problems such as the knapsack problem. The main feature of our method is its use of a binary decision diagram (BDD) or zero-suppressed BDD (ZDD) to represent the additional constraints. BDDs and ZDDs are data structures that represent a Boolean function as a compact graph. We first cast the additional constraints as a BDD or ZDD, then subject the resulting structure to a dynamic programming algorithm to solve the constraint-added problem in time linear to the number of BDD or ZDD nodes.



(a) BDD representing a Boolean function

x_1	x_2	x_3	$f(x_1, x_2, x_3)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

(b) Truth table representing the same Boolean function

Fig. 2. Representing a Boolean function using a binary decision diagram.

3. Binary decision diagrams (BDDs, ZDDs)

BDDs and ZDDs are data structures that represent a Boolean function as a graph. A Boolean function takes n input binary variables (variables that take either 0 or 1) and returns either 0 or 1. A BDD represents a Boolean function as a graph, as shown in **Fig. 2(a)**. There are several ways to represent a Boolean function other than with a BDD, such as a truth table (**Fig. 2(b)**) or a binary decision tree. The BDD differs from these representations in that it can represent an n-ary Boolean function as a BDD whose number of nodes is much smaller than 2^n; other representations require 2^n elements to represent the same Boolean function. The BDD also supports several operations that run in time proportional to the number of BDD nodes, and two BDDs can be subjected to Boolean operations to efficiently create another BDD.

A ZDD is a variant of a BDD and also represents a Boolean function as a graph. The ZDD differs from the BDD in that it can represent a family of sets as a DAG (directed acyclic graph) with fewer nodes than a BDD.

4. Combinatorial optimization using BDD and ZDD

We previously proposed some optimization algorithms that use BDDs and ZDDs [1, 2]. In this article, we discuss an algorithm that uses a ZDD to solve constraint-added variants of the 0-1 knapsack problem. The 0-1 knapsack problem is an optimization problem in which we are given n items that have their own costs and scores, and we must find the best subset of items that maximizes the sum of scores while keeping the sum of costs within the given threshold. A 0-1 knapsack problem can be solved efficiently by applying a dynamic programming algorithm.

Our algorithm can solve the problem made by adding constraints between variables to a 0-1 knapsack problem. As shown above, there is no efficient dynamic programming algorithm for ordinary 0-1 knapsack problems that contain additional constraints. However, if we represent such constraints by a ZDD, we can apply a dynamic programming algorithm that exploits the structure of the ZDD to solve constraint-added knapsack problems (**Fig. 3**). Our algorithm can solve a problem in $O(Z \times L)$ time, where Z is the number of ZDD nodes, and L is the threshold of total costs. We can efficiently solve such problems if the added constraints can be represented by a small ZDD.

5. Application to natural language processing

We applied our algorithm to text summarization, a common natural language processing task [2]. Text

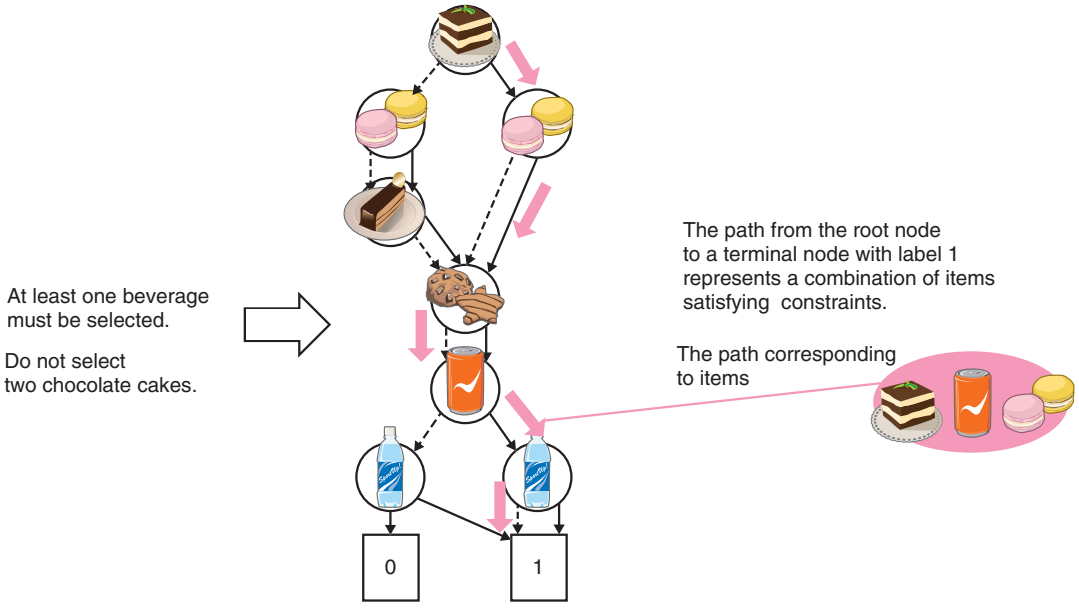


Fig. 3. Representation of constraints using ZDD.

summarization is the task of making a concise summary of an input document. Many text summarization approaches have been proposed, and one of the most popular approaches is to extract important sentences from the input document. Our research group proposed an extraction-based summarization method that first casts the input as a dependency structure tree that represents the dependency between clauses and then makes a summary by finding the optimal subtree [3]. Since this summarization method considers the dependency between clauses in making a summary, it can generate consistent summaries. Unfortunately, no efficient algorithm for solving this optimization problem has been developed until now.

This problem of finding the optimal subtree can be regarded as a 0-1 knapsack problem with the constraint that a solution must be a subtree of the input tree. This is a discrete constraint imposed on the relation between variables, so we represent it as a ZDD and solve the problem with our ZDD-based optimization method, which can solve problems up to 300 times faster than previous approaches. Furthermore, we prove that the number of ZDD nodes in the set of all subtrees of an input tree is bounded to $n \log n$, where n is the number of nodes of the input tree. This result suggests that our method can be applied to large problems.

6. Future work

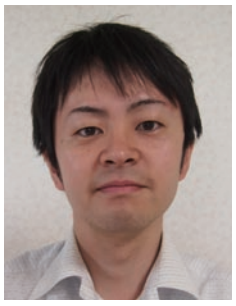
Since combinatorial optimization is relevant to many real world tasks, we believe our method can be used in fields other than natural language processing. We will continue to improve and analyze the performance of our algorithm in order to improve its effectiveness and efficiency.

References

[1] M. Nishino, N. Yasuda, S. Minato, and M. Nagata, "BDD-constrained Search: A Unified Approach to Constrained Shortest Path Problems," Proc. of AAAI-15 (Twenty-Ninth AAAI Conference on Artificial Intelligence), pp. 1219–1225, Austin, USA, Jan. 2015.

[2] M. Nishino, N. Yasuda, T. Hirao, S. Minato, and M. Nagata, "A Dynamic Programming Algorithm for Tree Trimming-based Text Summarization," Proc. of NAACL-HLT 2015 (2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies), pp. 462–471, Denver, USA, June 2015.

[3] T. Hirao, Y. Yoshida, M. Nishino, N. Yasuda, and M. Nagata, "Single-document Summarization as a Tree Knapsack Problem," Proc. of EMNLP 2013 (2013 Conference on Empirical Methods in Natural Language Processing), pp. 1515–1520, Seattle, USA, Oct. 2013.



Masaaki Nishino
Researcher, NTT Communication Science Laboratories.

He received his B.E., M.E., and Ph.D. in informatics from Kyoto University in 2006, 2008, and 2014. He joined NTT in 2008. His current research interests include natural language processing and combinatorial optimization.

Norihito Yasuda
Research Associate Professor, Graduate School of Information Science and Technology, Hokkaido University.

He received a B.A. in integrated human studies and an M.A. in human and environmental studies from Kyoto University in 1997 and 1999, and a D.Eng. in computational intelligence and system science from Tokyo Institute of Technology in 2011. He joined NTT in 1999. He is currently also a Research Associate Professor with the Graduate School of Information Science and Technology, Hokkaido University. His current research interests include discrete algorithms and natural language processing.

Tsutomu Hirao
Senior Research Scientist, NTT Communication Science Laboratories.

He received a B.E. from Kansai University in 1995, and an M.E. and Ph.D. in engineering from Nara Institute of Science and Technology in 1997 and 2002. His current research interests include natural language processing and machine learning.

Shin-ichi Minato
Professor, Graduate School of Information Science and Technology, Hokkaido University.

He received his B.E., M.E., and D.E. in information science from Kyoto University in 1988, 1990, and 1995. He worked for NTT from 1990 until 2004. He was a Visiting Scholar in the Computer Science Department of Stanford University in 1997. He joined Hokkaido University as an Associate Professor in 2004 and has been a full Professor since October 2010. He has served as the Research Director of the ERATO MINATO Discrete Structure Manipulation System Project, executed by the Japan Science and Technology Agency, since 2009. His research topics include efficient representations and manipulation algorithms for large-scale discrete structure data. He published "Binary Decision Diagrams and Applications for VLSI CAD" (Kluwer, 1995). He is a senior member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Information Processing Society of Japan (IPSJ), and a member of the Institute of Electrical and Electronics Engineers, and the Japanese Society for Artificial Intelligence (JSAI).

Masaaki Nagata
Senior Distinguished Researcher, Group Leader, NTT Communication Science Laboratories.

He received his B.E., M.E., and Ph.D. in information science from Kyoto University in 1985, 1987, and 1999. He joined NTT in 1987. He was with ATR Interpreting Telephony Research Laboratories from 1989 to 1993 and was a Visiting Researcher at AT&T Laboratories Research, New Jersey, USA, 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 IEICE, IPSJ, JSAI, the Association for Natural Language Processing, and the Association for Computational Linguistics.

Microscope Integrated with Optical Connector Cleaner for Cleaning and Inspecting Optical Fiber End-faces in a Single Operation

Yuichi Higuchi, Toru Miura, Koichi Hadama, and Joji Yamaguchi

Abstract

The end-faces of optical fibers must be kept clean because unclean end-faces can cause communication errors. When optical fibers are to be connected to each other, their end-faces are cleaned and inspected using two different devices in two separate time-consuming steps. First, a fiber cleaner is used to clean each end-face, and then a microscope is used to inspect each end-face. To simplify this process, we developed a device that integrates the cleaner and the microscope into a single tool, making it possible to perform both the cleaning and inspection in a single operation without having to change tools.

Keywords: optical connector, scope, cleaner

1. Introduction

Optical connectors are used for low-loss connection of optical fibers for equipment in telecommunication facilities. The end-faces of optical connectors are physically pressed together to achieve contact without gaps between the fiber cores. However, dust or other foreign material on the fiber end-faces can cause a gap or misalignment between the fibers, which can reduce the transmitted light or produce reflections and thereby degrade or disrupt communications. Moreover, foreign material can actually melt onto a fiber end-face, and the optical fiber itself can become fused if the connectors are used for high-powered light, for example, in optical line terminal video distribution or Raman amplification [1, 2].

To prevent these problems, connector end-faces should be inspected in accordance with relevant standards such as IEC[*1] Standard 61300-3-35 and ITU-T[*2] Recommendation L. 36. Such standards specify the permissible size and number of foreign particles in relation to the distance from the center of the fiber. Microscopic particles cannot be detected with the naked eye, so an optical connector microscope is used to photograph the fiber end-face, and the image is analyzed to determine if the inspection criteria are met. If the criteria are not met, a connector cleaner with a cleaning thread on the tip is used to clean the end-face. The end-face is then reexamined, and the inspection and cleaning processes are repeated until the criteria are met. The optical connector microscope and optical connector cleaner are separate, commercially available tools, and they are used individually in a process that involves repeated inspection and cleaning. This process can be tedious because each tool must often be switched and used a number of times.

To solve this problem and increase task efficiency,

*1 IEC: International Electrotechnical Commission
*2 ITU-T: Telecommunication Standardization Sector of International Telecommunication Union

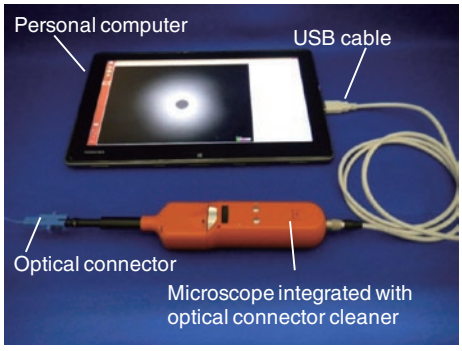


Fig. 1. Configuration of microscope integrated with optical connector cleaner.

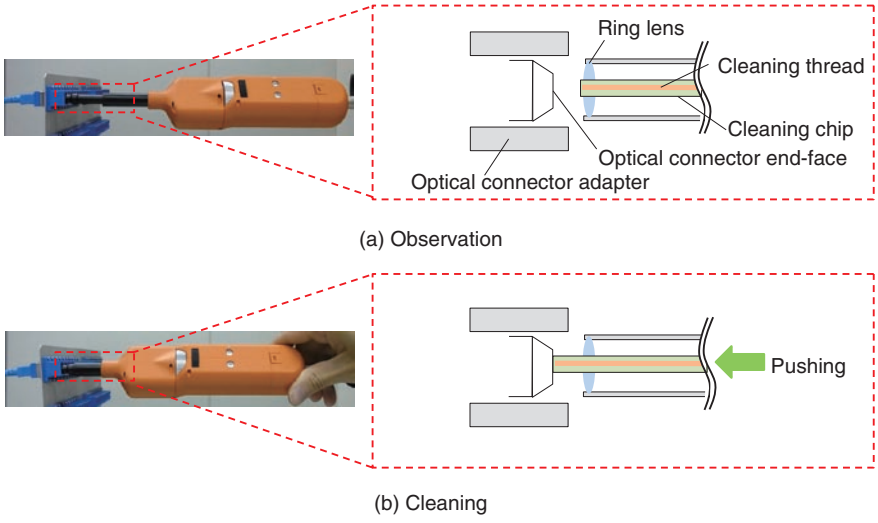


Fig. 2. Processes for (a) observation and (b) cleaning.

we propose integrating the optical connector cleaner and microscope into a single tool, making it possible to accomplish both tasks in a single operation.

2. Configuration and operation

The configuration of the microscope integrated with an optical connector cleaner is shown in **Fig. 1**. The integrated device is connected to a personal computer by a universal serial bus (USB) cable. The device consists of a light source for illuminating the connector, an image sensor for photographing the connector end-face, and imaging lenses. When the tip of the device is inserted into an optical adapter into which an optical connector plug has been inserted, an image of the connector end-face is formed on the image sensor. The image acquired by the sensor is transmitted via the USB cable to the computer for display.

The integrated cleaner/microscope device weighs 180 grams and measures 265 × 46 × 39 mm. It is easily used with one hand in the same way as existing optical connector cleaners and optical connector microscopes. The integrated device can be used with various types of connectors, including a single fiber coupling (SC) connector, miniature universal coupling (MU) connector, and a Lucent connector (LC), by changing the tip adapter.

The processes for using this device are illustrated in **Fig. 2**. The tip of the device is inserted into the adapter to observe the optical connector end-face (Fig. 2(a)). The cleaning step is then executed by pushing the device further into the connector adapter (Fig. 2(b)). The result of the cleaning operation can then be

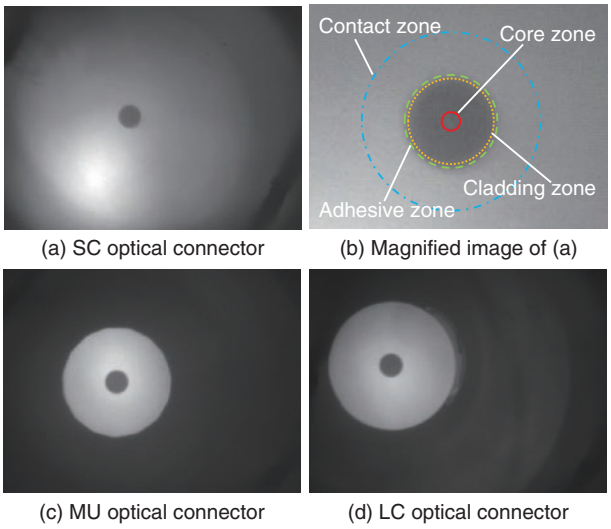


Fig. 3. Images of various optical connectors taken with integrated device.

observed by returning the device to the original observation position. Because it is not necessary to use two separate tools, the combined process is simpler and can be accomplished in a much shorter time.

The operation of this integrated device is simple. For observation, the device is equipped with a ring lens. The hole at the center of the lens accommodates a cleaning thread, which is used to clean the connector end-face. In the observation step, only the annular lens part is used. The connector end-face is illuminated, and the image of the connector end-face that is formed on the image sensor is acquired (Fig. 2(a)). In the cleaning step, pushing the device inward causes the cleaning thread to protrude from the ring lens and press against the optical connector, thus removing foreign material. This operation makes it possible to perform both the observation and cleaning tasks efficiently.

Optical systems that use a ring lens generally produce low-contrast images because the center part of the lens is missing. We designed the optical system so that it does not have this problem. The diameter of the hole at the center of the ring lens is determined by the size of the cleaning chip, and the shape of the adapter determines the aperture of the connector end-face. We took those conditions into account and optimized the lens size and the distance between the lens and the connector end-face. We thereby achieved optical characteristics that are about the same as those of a lens without a central hole.

The cleaner part of the system uses a cartridge-type cleaning thread that can be replaced when it is used up. The tip of the cartridge includes the lens, so the optical system is designed with a tolerance so that positional deviation due to replacement of the cleaner cartridge does not affect the optical characteristics.

3. Example of inspecting and cleaning optical connector end-face

Images of optical connector end-faces acquired with the integrated device are shown in **Fig. 3**, where (a) shows the end-face of an SC connector. The dark area in the center of the image is the optical fiber, and the white area around it is the protective ferrule. An enlarged view of the optical fiber area in Fig. 3(a) is shown in Fig. 3(b). For a single-mode fiber as specified in IEC 61300-3-35, the core zone extends from the center of the fiber to a diameter of 25 μm (indicated by the red line in Fig. 3(b)). The region from 25 to 120 μm from the center is the cladding zone (orange dotted line), from 120 to 130 μm is the adhesive zone (green dashed line), and from 130 to 250 μm is the contact zone (light blue dotted line). Each zone has a particular permissible size and quantity for foreign material. We can see from the image shown in Fig. 3(b) that the observable area extends fully up to the contact zone. Observed images for MU and LC connectors, for which the ferrule diameter is smaller and the adapter aperture is smaller, are presented in Figs. 3(c) and (d). Even for these connectors, the optical

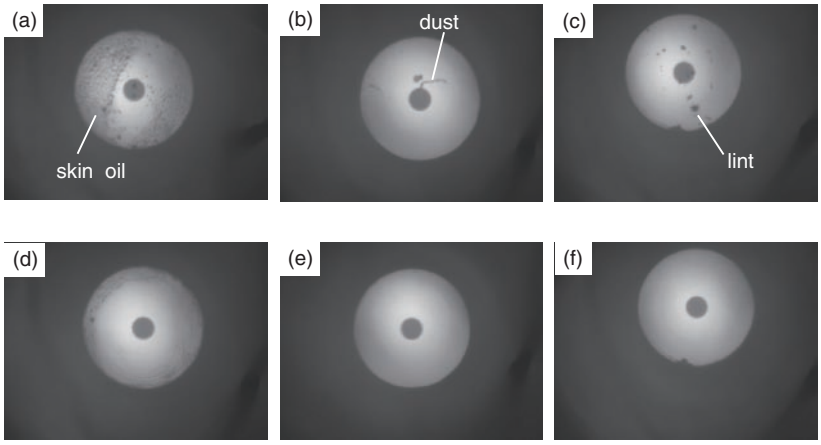


Fig. 4. Image of MU optical connector with (a) skin oil, (b) dust, (c) lint, (d) skin oil after cleaning, (e) dust after cleaning, and (f) lint after cleaning.

fiber and ferrule can be observed in the same way as for the SC connector, meaning that detailed images can be acquired regardless of the type of connector.

Example images before and after cleaning MU optical connector end-faces to which foreign material had adhered are shown in **Fig. 4**. The image in Fig. 4(a) clearly shows skin oil on the optical fiber ferrule. The ones in Figs. 4(b) and (c) show dust and lint on the optical fiber ferrule. The images presented in Figs. 4(d) to (f) show the connector end-faces after the cleaning was performed by pushing the integrated tool in and then returning it to the observation position. We can see from these figures that all of the different types of foreign materials were removed from around the optical fiber. These results demonstrate that the integrated inspection and cleaning tool operates effectively.

Whereas the conventional method for performing the two tasks requires an average of five steps (inspection, cleaning, inspection, cleaning, and inspection) and five tool insertions and removals, the proposed method requires one tool insertion and removal. As a result, using the integrated tool greatly shortens the total task time by eliminating most of the tool insertion and removal steps.

4. Performance evaluation

We evaluated the images acquired with this tool using the 1951 USAF[*3] test target, which is often used to measure the resolution of imaging systems (**Fig. 5**). The test image includes lines and spaces of various widths. Analysis of the contrast of the lines

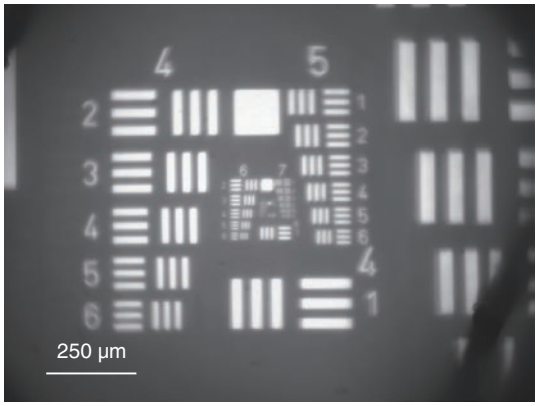


Fig. 5. Image of 1951 USAF test target taken with integrated device.

and spaces in an acquired image of the test pattern showed that the imaging system can resolve line widths of 2 µm or less. Given that the IEC 61300-3-35 standard requires the ability to identify defects as small as 2 µm, this integrated tool has sufficient optical resolution for practical application.

The cleaning performance of the integrated tool is presented in **Fig. 6**. The graph shows the test results for the number of times the cleaning operation had to be performed to meet the IEC 61300-3-35 standard. The cleaning operation was performed for an MU optical connector to which skin oil had adhered. For comparison, we conducted the same evaluation using

*3 USAF: United States Air Force

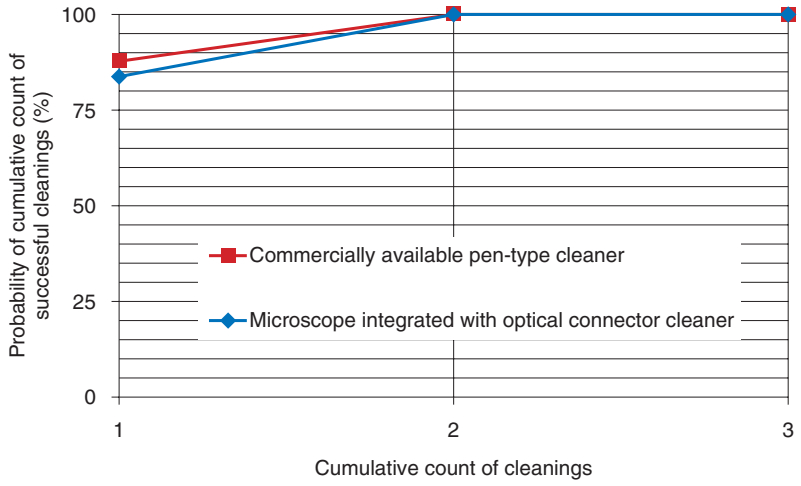


Fig. 6. Cleaning performance for MU optical connector using commercially available cleaner and integrated device.

a commercially available pen-type cleaner. We performed 50 trials for each test subject. A commercially available optical connector microscope and analysis software were used to evaluate whether the results of each trial met the criteria. The results showed that the integrated tool met the criteria with only two cleaning operations for every trial. Since the results were the same for the commercially available pen-type cleaner, we conclude that the integrated tool has the same cleaning performance as the commercially available cleaner.

5. Conclusion

The microscope integrated with an optical connector cleaner can reduce the time required for cleaning and inspecting the end-faces of optical connectors and thus reduce maintenance costs for telecommunication facilities. Evaluation of the imaging and cleaning performance of this tool demonstrated that it has sufficient optical resolution to meet the standards for optical connector inspection and has the same cleaning performance as existing optical connector cleaners.

References

[1] M. Ohmachi, M. Hosoda, M. Okada, M. Kihara, and M. Toyonaga, "A Fusion Fault of Optical Fiber Connector in Access Network," Proc. of the 2011 IEICE General Conference, B-10-25, Tokyo, Japan, Mar. 2011.

[2] A. Naka and T. Matsuda, "Operational Issues Facing Commercial Raman Amplifier System: Safety Measures and System Designs," Proc. of the Optical Fiber Communication Conference 2015, W3E.4, Los Angeles, USA, Mar. 2015.

Yuichi Higuchi
Researcher, Social Device Technology Laboratory, NTT Device Technology Laboratories.
He received a B.E. and M.E. in mechanical engineering from Kyoto University in 2006 and 2008. He joined NTT Microsystem Integration Laboratories in 2008. He is currently studying free-space optical devices for optical telecommunications and biosensing. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Japan Society for Precision Engineering (JSPE).

Toru Miura
Senior Engineer, NTT Advanced Technology Corporation.
He received a B.E. in optical engineering from the University of Electro-Communications, Tokyo, in 2000 and an M.E. and Ph.D. in optical engineering from Tokyo Institute of Technology in 2002 and 2005. He joined NTT Microsystem Integration Laboratories in 2005. Dr. Miura is currently in charge of the development of optical connectors and fiber microscopes. He is a member of the Japan Society of Applied Physics.

Koichi Hadama
Senior Research Engineer, Social Device Technology Laboratory, NTT Device Technology Laboratories.
He received a B.E. and M.S. in applied physics from the University of Tokyo in 1999 and 2001. He joined NTT Telecommunication Energy Laboratories in 2001. He is currently studying free-space optical modules for biosensing. He is a member of IEICE.

Joji Yamaguchi
Senior Research Engineer, Supervisor, NTT Device Innovation Center.
He received a B.E., M.E., and Ph.D. in mechanical engineering from Tokyo Institute of Technology in 1988, 1990, and 1993. In 1993, he joined NTT Interdisciplinary Research Laboratories, where he conducted research on optical cross-connect systems. Dr. Yamaguchi studied MEMS (microelectromechanical systems) control technology as a visiting researcher at the University of California, Berkeley, USA, from 2000 to 2001. He has recently been researching free-space optical devices for medical sensing. He is a member of the Japan Society of Mechanical Engineers and JSPE.

Trends in Standardization Activities in China

Daisuke Ikegami

Abstract

The China Communications Standards Association (CCSA) is the sole organization in charge of standardization in the Chinese telecommunications industry. This article introduces recent developments in the CCSA's standardization activities and explains the structure of the Chinese standardization system and trends in the telecommunications industry.

Keywords: standardization activities, CCSA, Chinese telecommunications

1. Introduction

The Chinese economy has grown by leaps and bounds in recent years. Development, however, has not stopped at the economy; the field of telecommunications has also seen remarkable growth. In particular, the infrastructure supporting telecommunications has been expanding rapidly. As of the end of 2014, the number of mobile phone subscriptions had reached nearly 1.3 billion, while the number of Internet users was 649 million.

In addition to the promotion of broadband, systematic measures are being devised and implemented in a variety of fields in China, following a unique Chinese model that entails considerable involvement by the government in the form of guidelines on the direction to take. In China, the central government lays out its policies every five years, determining the areas in which industrial development is to focus. During the period covered by the 12th Five-Year Guideline—from 2011 to 2015—seven industrial fields will be treated as key areas: energy conservation/environmental protection; next-generation information technology; biotechnology; the production of cutting-edge facilities; new energies; new materials; and cars employing new energy sources. At present, the 13th Five-Year Guideline, covering the next period (until 2020), is reportedly being drafted, and we believe that it broadly includes the direction to take in the field of telecommunications as well.

Starting this year in particular, the government began unveiling strategies for the use of telecommunications as a driver of economic growth. These include *Made in China 2025* and *Internet Plus*. Made in China 2025 identifies ten key areas of development: information technology; robotics/machine tools; aerospace; marine engineering; advanced railway facilities; energy conservation/energy-saving cars; power facilities; agricultural machinery; new materials; and biotechnologies/medical equipment. The objective is to make the manufacturing industry more efficient and to raise its standards by providing financial support and leveraging information and communication technology (ICT). The latter is expected to play a major role and to highlight ICT's increasing importance in the country.

2. China Communications Standards Association (CCSA)

2.1 Overview

The CCSA was founded in 2002 as the only standardization institution in charge of standards in the Chinese telecommunications industry [1]. The CCSA's membership is divided into full members, affiliate members, and observers, which encompass research institutions, telecommunications carriers, vendors, and universities. In recent years, membership has been on the rise, and as of 2014, there were 10 affiliate members and 32 observers in addition to

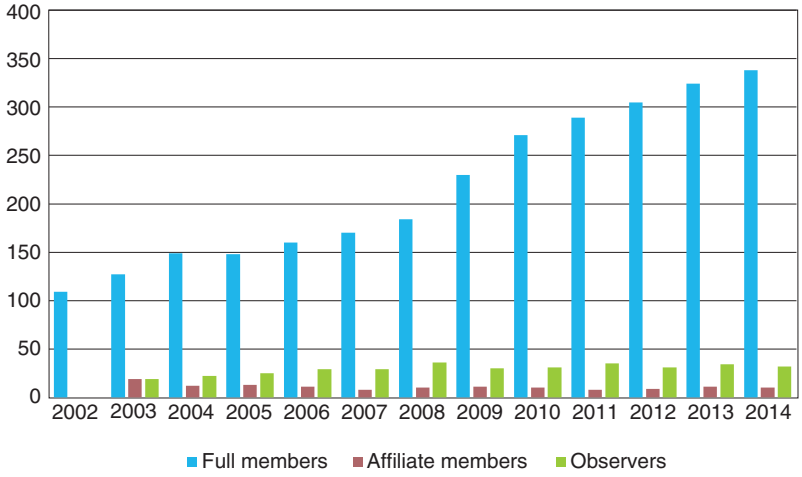


Fig. 1. Changes in CCSA membership since its establishment.

338 full members [2]. The changes in membership over the years are shown in **Fig. 1**.

In addition to activities involving carriers and vendors, CCSA's standardization activities revolve around the China Academy of Information and Communication Technology (CAICT)—China's only government-funded institution for research on telecommunications. CAICT provides support for government measures involving telecommunications, as well as providing consulting and certification services on standards. It is also actively involved in domestic standardization activities by Technical Committees (TCs) within the CCSA and in international standardization activities by the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T).

A list of the CCSA's TCs and Workgroups (WGs) for 2015 are listed in **Table 1**. The number of meetings held by each TC last year is indicated in **Fig. 2**, while the number of members attending is indicated in **Fig. 3**. The changes in the number of registered members of each TC since 2003 are shown in **Fig. 4**.

2.2 Main CCSA trends in 2014

In 2014, WG11 (peripheral facilities for wireless networks) was added to CCSA's TC5. This WG works on peripheral facilities that are not directly related to wireless network systems. This covers passive antennas for communication systems, traditional in-house sharing systems, broadband antenna products, passive components and their peripheral facilities, and the maintenance management of peripheral network facilities.

In 2014, 44 CCSA standards and 10 drafts for approval submitted by individual TCs were discussed throughout the CCSA; 40 CCSA standards and 8 draft proposals were approved. The total number of standards worked on by the CCSA in 2014 was 1647, including national standards, industry standards, CCSA standards, and research reports; of these, 465 standard drafts for approval were completed. This figure exceeded the beginning-of-year targets of 400 total drafts and 120 key standards. A breakdown of completed standards is given below.

- National standards: 9
- Industry standards: 379 (including 128 designated as key standards by the Ministry of Industry and Information Technology)
- CCSA standards: 17
- Research reports: 60

Such activity is remarkable even compared to that of our TTC (Telecommunication Technology Committee), now in its 30th year of foundation, which had a total of 837 enacted standards as of the end of 2014.

Of the standards completed by the CCSA, 168 national and industry standards were issued following a review by the Standardization Administration of the People's Republic of China and the Ministry of Industry and Information Technology. Reviews were also completed on 696 national and industry standards issued before 2009.

2.3 Slogan for standardization in 2014

CCSA's standardization work in 2014 was carried out in line with the slogan "Two levels, three areas."

Table 1. List of TCs and WGs set up within CCSA.

Study group		Subject studied
TC1 (IP and multimedia communications)	WG1	NW protocols and facilities
	WG2	IP services and applications
	WG3	Source encoding
	WG4	New technologies and international standards
	SWG2	IPTV
	SWG3	Future data networks (FDN)
TC3 (Networks and switching)	WG1	Networks in general
	WG2	Signaling protocols
	WG4	Services and applications
TC4 (Power supply for communications and operational environment of base stations)	WG1	Power supply for communications
	WG2	Communications room environment
TC5 (Wireless communications)	WG3	Broadband wireless access
	WG4	cdmaOne/CDMA2000
	WG5	3G network security and encryption
	WG6	Research on new frontier wireless technologies
	WG8	Frequencies
	WG9	TD-SCDMA/WCDMA
	WG10	Satellite/microwave communication
TC6 (Transmission networks and access networks)	WG1	Transmission networks
	WG2	Access networks and home networks
	WG3	Optical cable
	WG4	Optical devices
TC7 (Network management and operation support)	WG1	Wireless communication management
	WG2	Transmission, access and bearer network management
	WG3	ICT service management and operation
TC8 (Network and information security)	WG1	Wired network security
	WG2	Wireless network security
	WG3	Security management
	WG4	Security infrastructure
TC9 (Electromagnetic environment and protection)	WG1	Electromagnetic environment of telecommunications facilities
	WG2	Measures against thunderstorm damage and environmental adaptability of telecommunications systems
	WG3	Electromagnetic radiation and safety
TC10 (Ubiquitous networks)	WG1	General
	WG2	Applications
	WG3	Networks
	WG4	Sensing/development
TC11 (Mobile Internet application and terminals)	WG1	General
	WG2	Service platforms and their application
	WG3	Terminals

CDMA: code division multiple access
NW: network
IP: Internet protocol
IPTV: IP television
SWG: Special Workgroup
TD-SCDMA: time-division synchronous code division multiple access
WCDMA: wideband code division multiple access

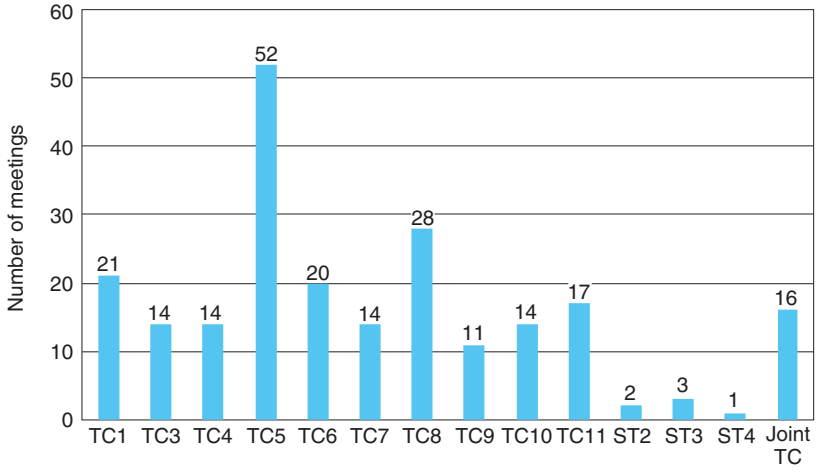


Fig. 2. Number of meetings held by each TC in 2014.

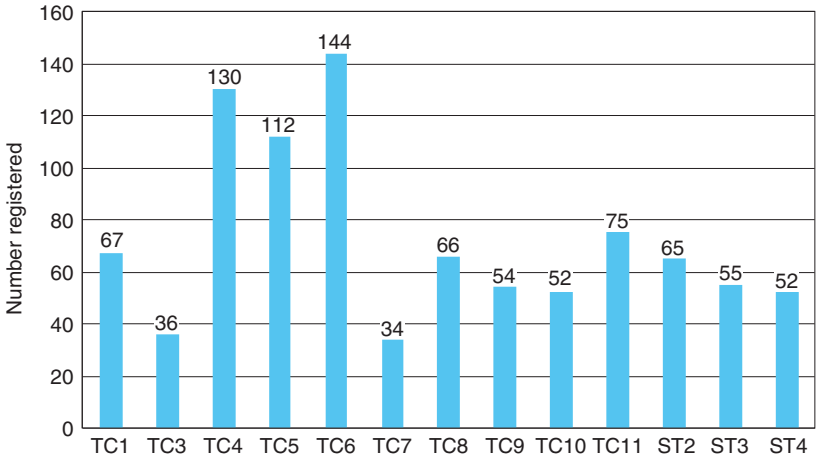


Fig. 3. Number of registered members of each TC in 2014.

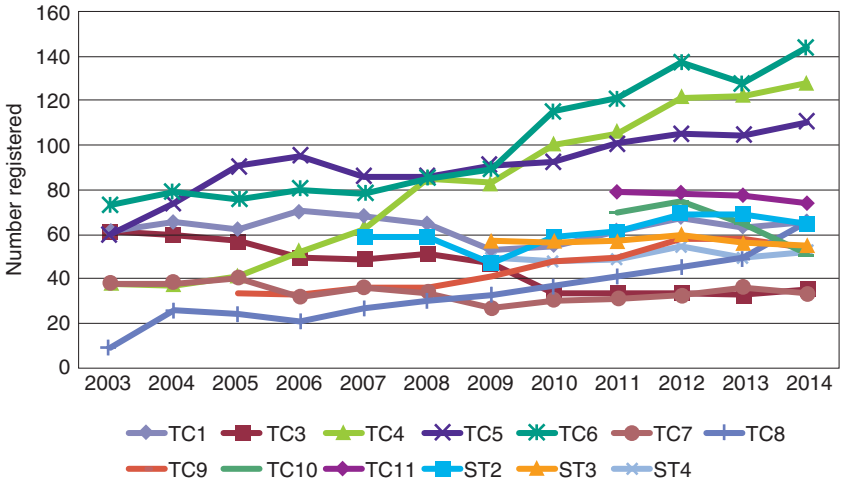


Fig. 4. Changes in number of registered members for each TC since CCSA's establishment.

Here, "two levels" means to conduct standardization under different rules in China and at an international level in view of the need to adapt to differences in the nature, areas, and direction of standardization. The aim is not consistency with international standards, but rather, flexible standardization to suit circumstances in different areas within China.

The expression "three levels" indicates the government's intent to place particular focus on the following three areas.

(1) Level 1: Standardization to meet urgent needs for purposes such as industrial development and government regulations
(2) Level 2: Software-defined networking (SDN), cloud computing, and other new areas of technology
(3) Level 3: 5G and other areas of technology with future potential

3. CCSA standardization in 2014

3.1 Key areas of domestic Chinese standards

In 2014, the CCSA worked on three key areas of standardization: standards required by governmental sector regulations and public services; standards issued in a timely fashion to meet the needs of industrial development; and support for industries and member companies to begin new standardization

work. The CCSA indicates the specific nature of standardization initiatives undertaken on a priority basis in each area as follows.

The five key areas of standardization for standards required by governmental sector regulations and public services are:

(1) The standard series relating to Security Capabilities for Smart Mobile Terminals
(2) The standard series relating to Protection of Personal Information in Telecommunications and Internet Services
(3) Standards for domain names in Chinese
(4) Technical Requirements for the Mutual Exchange of Information in Smart Cities
(5) The standard series relating to electronic IDs (identification) in Network Spaces

The six key areas of standardization for standards issued in a timely fashion to meet the needs of industrial development are:

(1) The standard series relating to LTE (Long Term Evolution) Multi-Module Single-Card Terminals
(2) The development of a standard system for managing LTE networks
(3) Packet expansion-type OTN (optical transport network) (facility function model adopted by ITU-T as well)
(4) IPv6 (Internet protocol version 6) address management
(5) Remote positioning for electric bikes
(6) High-voltage DC (direct current) 240 V/336 V power supply systems

The five key areas of standardization for support for industries and member companies to begin new standardization work are:

(1) ITU-T's enactment of PTN (packet transport network) standards enacted by the CCSA
(2) Strategically breaking through barriers of patents for video encoding held by foreign companies
(3) Smart ODNs (optical distribution networks)
(4) OTA (over-the-air) testing standards for LTE mobile terminals
(5) Industry standards for adult head models with numerical evaluation of short-range electromagnetic radiation, in which China has intellectual property rights to the developed technology

3.2 Key areas of international standardization

The national government is also focusing its efforts on international standardization and is bolstering its support for the appointment of executives of international standardization bodies and article contributions. China is now an important player in the field of international standardization and makes a remarkable contribution of over 7000 articles each year to ITU, the 3rd Generation Partnership Project (3GPP), IEEE (Institute of Electrical and Electronics Engineers), the Internet Engineering Task Force (IETF), and other associations. In this context, China is focusing on SDN, LTE, and cloud computing, and is emphasizing the results through CCSA's reports, among other things. For example, results obtained in the field of SDN include the creation of Y.2301, a standard for smart communication networks in which China is a leader, and the creation of working items for SDN-based smart pipe architectures.

In the field of LTE, China has formulated an international proposal centered around air interfaces that meet API (application programming interface) technical requirements (the first step) for the use of broadband trunking communications systems (B-TrunC) employing LTE technology—the Chinese industry's standard. In November, this was adopted as proposal ITU-R (Radiocommunication Sector) M.2009, and was reportedly established as the broadband trunking air interface for PPDR (Public Protection and Disaster Recovery) in ITU-R Recommendations.

Moreover, in the field of cloud computing, China is leading ITU-T's review on cloud computing frameworks in accordance with the cloud computing standard created by the CCSA. It has completed four standards, which include high-level and infrastructure requirements as well as resource management frameworks (Y.3501, Y.3510, Y.3520, and X.1601). Twelve standardization projects in other areas of cloud computing and big data are also reportedly in progress.

China is also working with standardization bodies other than ITU. In 3GPP, articles contributed by China in the field of wireless technology now account for over 25% of the total, and about one-third of the 3GPP executives are Chinese. China has also made its presence felt within the IETF, and Chinese companies have completed 195 Requests for Comments (RFCs) to date. This figure accounts for about 2.71% of all RFCs, placing China ninth in international rankings. The WG on ACTN (Abstraction and Control of Transport Networks) is currently being led by China, according to reports.

Furthermore, the Chinese government has developed a support framework for international standardization work and is providing such support proactively in the form of subsidies. In particular, in the

Table 2. Key areas of work engaged in by each TC in 2014.

TC name	Key areas of work in 2014
TC1	IPv6-based next-generation Internet
	Cloud computing
	FDNs
	SDN and NFV (network function virtualization)
	Internet work
	New technologies in the IP bearer network area
	Signal coding and meta-data
	Big data
	Removal of information barriers
TC3	Smart communication networks
	Unified IMS (IP multimedia subsystem)
	SDN/NFV
	RCS (Revision Control System) operations
	Open APIs for operational capabilities based on REST (Representational State Transfer)
TC4	240 V/360 V high-voltage DC power supply
	Energy efficiency at datacenters
	Renewable energy
TC5	LTE terminals
	LTE broadband trunk B-TrunC
	Public wireless area networks
	Research on frequencies
	Peripheral equipment and facilities
	Satellite/digital microwaves, security
TC6	PTN (packet transport network) series standards
	Standards of packet expansion-type OTN (optical transport network) series
	Standards for the smart ODN (optical distribution network) series
	400G/1T high-speed transmission technology
	OTN technologies over 100G
	SDTN (software defined optical transport network)
	TWDM-PON (time-wavelength division multiplexing passive optical network)
TC7	LTE network management
	5G network management
	Big data management
TC8*	Information security for mobile Internet
	Security for new mobile Internet businesses
	Security management for mobile Internet
	Cloud computing security
	Electronic IDs for domain spaces
	* Released to joint and foreign-invested companies also from this year onwards.
TC9	Measurement of performance of mobile terminal antennas
	Electromagnetic compatibility and protection
	Lightning surge protection
	Electromagnetic radiation
TC10	M2M (machine to machine)
	Remote positioning services for electric bikes
	Smart cities
	Vehicle-to-vehicle networks

field of telecommunications, the CCSA, entrusted by the development department of the Ministry of Industry and Information Technology, accepts applications for subsidies in aid of international standardization projects in the telecommunications industry. In the first quarter of 2014, the amount of 420,000 RMB was provided for 29 projects, while 550,000 RMB was provided for 30 projects in the second quarter. The amount for 26 projects during the third quarter was under review as of December 2014. With the aim of providing subsidies to help with the expenses of drafting international standards, the CCSA also enacted the CCSA Administrative Measures for Subsidies for International Standardization in the Communications Industry in 2014, accentuating the fact that the entire country's resources are being tapped into in order to promote international standardization.

3.3 Key areas of work by each TC in 2014

Additional areas worked on by individual TCs in 2014 are listed in **Table 2**.

At the NTT Beijing Representative Office, we are participating in WG1 of CCSA's TC10 as observers in order to gather information. In 2014, TC10 focused on standardization in four areas: machine to machine (M2M), remote positioning services for electric bikes, smart cities, and vehicle-to-vehicle networks.

In the field of M2M, TC10 created ratified drafts of industry standards concerning "Technical Requirements for Service Capabilities of M2M Terminals" and "Technical Requirements for M2M Service Platforms." It simultaneously developed a more in-depth discussion of the existing "General Technical Requirements of M2M Service" and "Requirements of M2M Communication Protocols."

TC10's work in 2014 emphasized in particular the

industry standards concerning positioning services for electric bicycles. These standards consist of service platforms, technical requirements for positioning services, and measurement methods for peripheral components. In establishing these standards under the lead of the Ministry of Industry and Information Technology and the Ministry of Public Security of the PRC, the CCSA and public security system staff engaged in cross-departmental cooperation, which resulted in the joint creation of a standard series on remote positioning services for electric bicycles by local government bodies and corporations. These standards make it possible to reduce economic losses from theft and other factors, and their commercialization is already underway in more than ten provinces and cities. As of September 2014, 2 million electric bicycles had been sold, with sales across the industry reaching 520 million RMB, and profits reported at 130 million RMB.

Standardization in the field of smart cities is underway in response to the demand generated by the construction of smart cities in China. The standardization review has encompassed smart city standard systems, public support platforms, information-sharing technologies, open data requirements, construction management, and service models. This has resulted in the issuance of the industry standard "Technical Requirements for the Mutual Exchange of Information in Smart Cities."

In the field of vehicle-to-vehicle networks, TC10 has been conducting a review since 2011 and has produced numerous CCSA standards, which include: "Service Requirements and General Framework for Vehicle-to-Vehicle Informatization Employing Ubiquitous Networks"; "General Framework for Smart Transportation Systems Supported by Communications Networks"; and "General Technical Requirements for Vehicle-to-Vehicle Networks." TC10 has also completed a draft for approval on the industry standard "Technical Requirements for Public Communications Network-Based In-Vehicle Gateways," and has completed a series of standards.

Additionally, TC10's WG1 has systematized a series of nine standardization documents on smart cities and is pushing forward with a review. WG1 is also working on standardization in relation to healthcare services employing mobile Internet and the Internet of Things (IoT) and has begun standardization efforts on "IoT-based Mobile Health (Needs)" and "Electronic Health Service Categories." The main targets of these standards are the needs faced in providing healthcare services through mobile Internet and use case categories; in the future, however, the review is expected to move on to more specific models, and we may need to keep track of future trends.

4. Conclusion

This article has provided a summary of the CCSA—the only standardization body for telecommunications in China—and of the main areas of standardization that the CCSA worked on in 2014. China is seeking to use ICT as a driver for its industrial development. We expect that it will not only push forward with domestic standardization but will also intensify its efforts in international standardization. In particular, we will need to keep an eye on trends in the areas on which the government is focusing. At the NTT Beijing Representative Office, we will continue to gather information on standardization trends in the Chinese telecommunications industry on behalf of the NTT Group and to pursue/promote new cooperative relationships with China.

References

[1] Website of CCSA, http://www.ccsa.org.cn/
[2] CCSA, "State of Organizational Development and Technical Activities in FY 2014," Dec. 2014.

Daisuke Ikegami
Associate Manager, Beijing Representative Office, Global Business Office, NTT*.
He received his B.E., M.E., and Ph.D. from Waseda University, Tokyo, in 2002, 2004, and 2007. Since joining NTT Service Integration Laboratories in 2007, he has been engaged in research on IP service network engineering. He is currently engaged in promotional activities for research and development technologies at NTT Beijing Representative Office. He received the Young Engineer Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2010. He is a member of IEICE.
* He is currently with NTT Network Technology Laboratories.

Event Report: NTT Communication Science Laboratories Open House 2015

Tessei Kobayashi, Hideki Sakurada, Sanae Fujita, Shiro Kumano, and Nobutaka Ito

Abstract

NTT Communication Science Laboratories Open House 2015 was held in Keihanna Science City, Kyoto, on June 4 and 5, 2015. Over 1200 visitors attended the event and enjoyed 5 talks and 30 exhibits focusing on our latest findings and activities in the fields of information and human sciences.

Keywords: information science, big data, human science

1. Overview

At NTT Communication Science Laboratories (NTT CS Labs) located in Seika-cho, Kyoto, and Atsugi City, Kanagawa, we aim to build new technical infrastructures connecting *people* and *information* by creating innovative technologies and discovering new principles. The targets of NTT CS Labs are the most fundamental research topics in the fields of information and human sciences.

NTT CS Labs Open House has been held annually for the purpose of introducing our innovative technologies and new findings in basic research to many visitors from the NTT Group, other industries, universities, and research institutions who are engaged in research, development, business, and education.

This year, Open House was held at the NTT Keihanna Building in Kyoto on June 4 and 5, and over 1200 visitors attended it over the two days. We prepared many hands-on exhibits to allow visitors to intuitively understand our latest research results and to share a vision of the future where new products based on the research results are widely used. We also organized an invited talk on the themes of human interfaces and robotics. This article summarizes the event's research talks and exhibits.

2. Keynote speech

Open House started with a speech by Director of NTT CS Labs, Eisaku Maeda, entitled "Embracing information science and technology—Decoding, exploring and designing the world" [1] (**Photo 1**).

In this talk, he mentioned that the era in which people are confronted with machines (e.g., computers and artificial intelligence) is coming to an end, and we humans will need to internalize information science and technology as part of ourselves. He also stressed that many researchers need to acquire the

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

Photo 2. Research talk by Naonori Ueda.

Photo 3. Research talk by Masaya Murata.

ability to decode, explore, and design the entire world including us (i.e., humans). Moreover, he argued that while bearing in mind the drastic changes in the information environment that we have experienced in the first 15 years of the 21 century, we must think about what concepts should make up basic research that will form the compass of the future as we envision the year 2030, 15 years from now. He proposed a new direction on information science and technology, and stressed the importance of three new frameworks: *Measuring to Understanding*, *Analysis to Exploration*, and *Implementation to Design*.

3. Research talks

Three research talks were held that highlighted the recent significant findings and activities of NTT CS Labs. Each talk provided an overview of the research field and introduced the latest research results. All of the talks were very well received.

(1) "'When', 'where', 'what', and 'how'?—Spatio-temporal multidimensional data analysis for IoT* big data," by Naonori Ueda, Machine Learning and Data Science Center [2]. He introduced the basic concepts and some examples of spatio-temporal multidimensional collective data analysis that can predict when, where, and what in real time and provide feedback to social systems (**Photo 2**).

(2) "What is this? Who is it? Where am I?—Recent advances in real-world media search technology and future," by Masaya Murata, Media Information Laboratory [3]. He summarized previous work in the field of media search and then introduced recent findings and activities on his project on instance search technology (**Photo 3**).

(3) "Hidden hearing processes unveiled—Exploring auditory mechanisms by biological measures," by Shigeto Furukawa, Human Information Science Laboratory [4]. He explained current findings and the direction of auditory mechanism research and then introduced innovative methods for measuring human auditory ability by using multiple biological responses to visual as well as auditory stimuli (**Photo 4**).

Photo 4. Research talk by Shigeto Furukawa.

4. Research exhibits

Open House featured 30 exhibits displaying our latest research activities. The exhibits were classified into four categories: (1) big data science, (2) computer

* IoT: Internet of Things

science, (3) media intelligence, and (4) communication and human science. Each exhibit was housed in a booth and employed 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 in each category. More details including the names of researchers associated with each exhibit can be found on the Open House website [5, 6].

4.1 Big data science

(1) Automatic tailor-made data analysis
—Generating probabilistic models using structure information—

(2) Finding various factors hidden in data
—Advanced and fast high-dimensional multiple factorization—

(3) Infinite data analysis beyond big data
—Stochastic process models for infinite-dimensional matrices—

(4) Fast graph analysis by efficient CPU utilization
—Scalable parallel graph processing by reordering—

(5) Efficient knowledge discovery from large-scale graph
—Efficient structure mining for large-scale graph—

(6) Agile environmental sensing
—CILIX: a virtual machine for wireless sensor network applications—

(7) Satisfying visitors with collective navigation while predicting people flow
—Spatio-temporal data analysis for controlling massive people flow—

(8) Traffic flow aggregation for traffic engineering
—Classifying flows based on traffic variation pattern—

(9) ESKORT: mining expert knowledge from trouble ticket
—Automatic workflow extraction from unstructured texts for network operation—

4.2 Computer science

(10) Can you beat TOROBO-kun?
—Evaluating the naturalness of sentences using language models—

(11) "Wanwan" is easier to learn than "inu"
—Exploring differences in learnability of IDS and ADS words in children—

(12) Better Japanese understanding helps better translation
—Pre-ordering machine translation by deep syntactic analysis—

(13) Finding the best combination within budget
—Combinatorial optimization using binary decision diagram—

(14) Quantum computer using noisy environment
—Identification of untouchable quantum system—

(15) Guaranteeing randomness
—A new concept toward reliable physical random bit generation—

4.3 Media intelligence

(16) Analyzing, synthesizing and converting speech prosody
—Generative modeling of voice fundamental frequency contours—

(17) Understanding by capturing
—Simultaneous multi-biosignal sensing by visible light communication— (**Photo 5**)

(18) Connecting what you see to the information world
—Real-world information retrieval by media search technology—

(19) Recovering perfectly-sounding speech by denoising
—High-quality speech enhancement using vast amounts of examples—

(20) Extracting essential information from sounds
—Advances in distant speech recognition by deep learning—

(21) Measurement of fluorescence by 9-eye camera
—"Shitsukan" reproduction using reflection and fluorescent of objects—

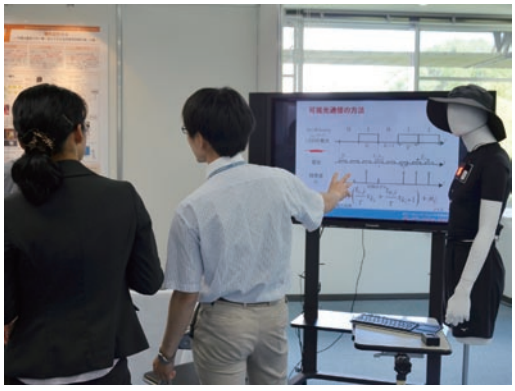


Photo 5. Exhibit: "Understanding by capturing."

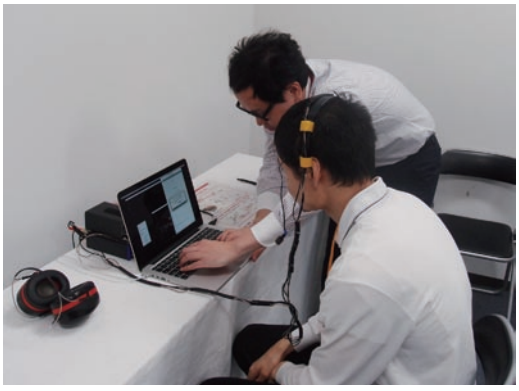


Photo 6. Exhibit: "We can see how you hear."

4.4 Communication and human science

(22) ICT support for mental health care
—Understanding the burden of family caregivers of a depressed individual—

(23) Who will start speaking next and when?
—Predicting next speaker and timing using gaze and breathing—

(24) Human thermal sense doesn't work as a thermometer
—Exploring mechanisms of human thermal information processing—

(25) New text communication using finger tracing
—Creating dynamic texts using "Yu bi Yomu" applications—

(26) Hen-Gen-Tou (Deformation Lamps)
—Amazing illumination to make static objects dynamic—

(27) Where is my hand?
—Bodily response induced by an illusion in the sense of body—

(28) I can speak like a native
—Speaking rhythm conversion rules by English speech corpus—

(29) Getting the knack of action and player's condition
—Body / mind reading & feedback system for sports—

(30) We can see how you hear
—Biological measures for auditory experience— (**Photo 6**)

5. Invited talk

This year, we invited Prof. Masahiko Inami from Keio University's Graduate School of Media Design. He gave a talk entitled "Initial step towards augmented human." He talked about his work on the augmentation of human senses and perception by technology. One of his well-known research topics is an optical camouflage cloak, which enables a masked object to be observed as if it were virtually transparent. He demonstrated his efforts to make a car's backseat virtually transparent and allow drivers to see the road behind the backseat. He has been developing the concept of augmented humans, and recently founded the Superhuman Sports Society, which is creating and promoting sports played using augmented human technologies.

6. Promotion using the web

To inform many people about our research activities, we created both Japanese [5] and English websites [6] for Open House 2015, which included a booklet, exhibition posters, and reference information (**Fig. 1**). We also uploaded photos of exhibition halls and videos of the three research talks and the Director's keynote speech. Furthermore, we tweeted about the exhibition events via Twitter [7] and Facebook [8] through NTT's Public Relations Office, and we provided an online research lecture via the video streaming platform "niconico Live" offered by

Fig. 1. English website of NTT CS Labs Open House 2015 [6].

Dwango Co., Ltd. [9]. In this lecture, Hirokazu Kameoka (Media Information Laboratory) talked about analyzing, synthesizing, and converting speech prosody, and he received many online comments from about 500 anonymous web users [10]. Thus, NTT CS Labs is continuously trying to improve the ways of disseminating our research activities and results.

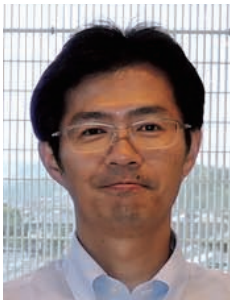
7. Concluding remarks

Following the success of last year's event, many visitors came to NTT CS Labs Open House 2015 and engaged in lively discussions on the research exhibits and talks. All of the advice and comments from visitors are very valuable to the members of NTT CS Labs, and they help to support and speed up our research activities. Therefore, we would like to offer our sincere thanks to all of the visitors and participants who attended this event.

References

[1] Video streaming site of keynote speech by Eisaku Maeda (in Japanese).
http://www.kecl.ntt.co.jp/openhouse/2015/talk/director/index.html
[2] Video streaming site of research talk by Naonori Ueda (in Japanese).
http://www.kecl.ntt.co.jp/openhouse/2015/talk/research1/index.html
[3] Video streaming site of research talk by Masaya Murata (in Japanese).
http://www.kecl.ntt.co.jp/openhouse/2015/talk/research2/index.html
[4] Video streaming site of research talk by Shigeto Furukawa (in Japanese).
http://www.kecl.ntt.co.jp/openhouse/2015/talk/research3/index.html
[5] Japanese website of NTT Communication Science Laboratories Open House 2015.
http://www.kecl.ntt.co.jp/openhouse/2015/index.html
[6] English website of NTT Communication Science Laboratories Open House 2015.
http://www.kecl.ntt.co.jp/openhouse/2015/index_en.html
[7] Twitter account by NTT's Public Relations Office.
https://twitter.com/NTTPR
[8] Facebook page by NTT's Public Relations Office.
https://www.facebook.com/NTTgroup?fref=ts
[9] Niconico Live by Dwango.
http://live.nicovideo.jp/
[10] Online lecture by Hirokazu Kameoka (in Japanese).
http://live.nicovideo.jp/watch/lv222082115

Tessei Kobayashi
Senior Research Scientist (Distinguished Researcher), Interaction Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.
He received a Ph.D. in psychology from the University of Tokyo in 2004. He is currently engaged in research on child language development, especially vocabulary spurts and syntactic bootstrapping.



Hideki Sakurada
Senior Research Scientist, Computing Theory Research Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received an M.Sc. in information science from Kyoto University in 1999. He is currently researching formal methods of network security, especially analysis of cryptographic protocols.

Sanae Fujita
Research Scientist, Linguistic Intelligence Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.
She received a Ph.D. in engineering from Nara Institute of Science and Technology in 2009. She is currently engaged in research on natural language processing, especially word disambiguation and text readability estimation.

Shiro Kumano
Research Scientist, Sensory Resonance Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories.
He received a Ph.D. in information science and technology from the University of Tokyo in 2009. He is currently conducting research on affective computing and computer vision, especially automatic analysis of affect/behavior in conversation.

Nobutaka Ito
Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received a Ph.D. in information science and technology from the University of Tokyo in 2012. He is currently researching audio signal processing, especially speech enhancement using a microphone array.

New NTT Colleagues
—We welcome our newcomers to the NTT Group

This is a corner of the NTT Technical Review where we introduce our new affiliate companies.

iPay88
Online payment service and payment solution provider in Malaysia; established in 2006; headquartered in Kuala Lumpur

Founded in 2006, iPay88 provides online payment services and payment solutions for merchants and financial institutions in Malaysia. The company has established a strong position with a leading share based on transaction volume and the number of established e-commerce merchants using iPay88's services already operating in Malaysia. In addition, since consumers in the Asia-Pacific region tend to use various methods of payment, including not only credit cards but also direct debit and prepaid cards, iPay88's comprehensive payment options enhance merchants' competitiveness in capturing both local and overseas markets.

In September 2015, NTT DATA Corporation, through its subsidiary NTT DATA Asia Pacific Pte. Ltd, completed its acquisition of iPay88 Sdn. Bhd. For NTT DATA, this transaction opens a gateway to establishing a firm foothold in the online payment business in the growing Asia-Pacific market. For further information about iPay88, please visit http://www.ipay88.com/

Contact:
Public Relations Department
NTT DATA Corporation
http://www.nttdata.com/global/en/news-center/others/2015/090400.html

External Awards

JSAI Incentive Award
Winner: Hiroaki Sugiyama and Toyomi Meguro, NTT Communication Science Laboratories; and Ryuichiro Higashinaka, NTT Communication Science Laboratories/NTT Media Intelligence Laboratories
Date: June 12, 2015
Organization: The Japanese Society of Artificial Intelligence (JSAI)

For "Experimental Analysis for Automatic Evaluation of Open-domain Conversational Systems based on Large-scale Multi-references."

The evaluation of conversational systems that chat with people remains an open-problem. Some studies have evaluated them by hand with ordinal scales such as the Likert scale. One limitation with this approach is that we cannot use the previously evaluated values since the ordinal scales are not consistent across all of the evaluations. This makes it difficult to compare proposed and previous systems since we have to implement the previous systems and simultaneously evaluate them. We propose an automatic evaluation method for conversational systems that evaluates sentences generated by systems on the basis of similarities that are calculated with many reference sentences and their annotated evaluation values.
Published as: H. Sugiyama, T. Meguro, and R. Higashinaka, "Experimental Analysis for Automatic Evaluation of Open-domain Conversational Systems based on Large-scale Multi-references," SIG-SLUD, Vol. B4, No. 01, pp. 1–6, Sept. 2014.

IEICE ISS Young Researcher's Award in Speech Field
Winner: Ryo Masumura, NTT Media Intelligence Laboratories
Date: August 21, 2015
Organization: Institute of Electronics, Information and Communication Engineers (IEICE) Speech Committee

For "Investigation of Combining Multiple Language Modeling Techniques in Japanese Spontaneous Speech Recognition."

Recent large vocabulary speech recognition systems consist of two statistical models, the acoustic and language models. In acoustic modeling, deep neural networks have realized a breakthrough, and significant performance improvements have been achieved. On the other hand, in language modeling, there have not been any reports of comparable improvements. Although it is clear that recent practical language models have several problems such as "locality," "task dependency" and "data sparseness," we cannot obtain significant performance improvements by solving these problems separately. In this paper, we try to use various language modeling techniques simultaneously to cover the entire problem. Our investigation was conducted by dividing language modeling techniques into three viewpoints, "one pass decoding," "unsupervised adaptation" and "rescoring."
Published as: R. Masumura, T. Asami, T. Oba, H. Masataki, and S. Sakauchi, "Investigation of Combining Multiple Language Modeling Techniques in Japanese Spontaneous Speech Recognition," IEICE Tech. Rep., Vol. 114, No. 151, SP2014-63, pp. 1–6, Jul. 2014.

IEEJ Excellent Presentation Award
Winner: Takuya Hoshi, NTT Device Technology Laboratories
Date: August 26, 2015
Organization: The Institute of Electrical Engineers of Japan (IEEJ)

For "Impact of Strained GaAs Spacer between InP Emitter and $GaAs_{1-y}Sb_y$ Base on Structural Properties and Electrical Characteristics of MOCVD-grown InP/$GaAs_{1-y}Sb_y$/InP DHBTs."

Novel InP/$GaAs_{1-y}Sb_y$/InP double-heterojunction bipolar transistors (HBTs) with a GaAs spacer between the InP emitter and $GaAs_{1-y}Sb_y$ base layer were grown by the metalorganic chemical vapor deposition method in order to simplify the switching sequence for forming a high-quality InP-emitter/$GaAs_{1-y}Sb_y$-base interface. The insertion of the GaAs spacer is a good way to obtain a high-quality E-B interface with a simple precursor-supply sequence and thereby HBTs with both high-current gain and reasonably high RF performance.
Published as: T. Hoshi, N. Kashio, H. Sugiyama, H. Yokoyama, K. Kurishima, M. Ida, H. Matsuzaki, and M. Kohtoku, "Impact of Strained GaAs Spacer between InP Emitter and $GaAs_{1-y}Sb_y$ Base on Structural Properties and Electrical Characteristics of MOCVD-grown InP/$GaAs_{1-y}Sb_y$/InP DHBTs," Journal of Crystal Growth, Vol. 395, pp. 31–37, Jun. 2014.

IEICE Fellow
Winner: Takehiro Moriya, NTT Communication Science Laboratories
Date: September 9, 2015
Organization: Institute of Electronics, Information and Communication Engineers (IEICE)

For contributions to research, development and standardization of high-compression coding schemes for speech and audio signals.

IEICE Fellow
Winner: Masahito Tomizawa, NTT Network Innovation Laboratories
Date: September 9, 2015
Organization: Institute of Electronics, Information and Communication Engineers (IEICE)

For contributions to research, development and standardization of technologies for large capacity optical transport networks.

JCSS Encouragement Paper Prize
Winner: Noriko Shingaki, Seijo University; Miki Kitabata, NTT Network Innovation Laboratories; Hiroto Matsuoka, NTT Device Innovation Center; Toshihiro Takada, NTT Communication Science Laboratories; Akiko Orito, J. F. Oberlin University; Yuko Kato, CDI; Yukie Tsuzuki, Seijo University; and Tatsuo Owada, NTT Resonant Inc.
Date: September 19, 2015
Organization: Japanese Cognitive Science Society (JCSS)

For "Accuracy and Distortions of Personal Memories ("Omoide") Saved in a Nine-year Time Capsule."

How should we save our memories? Many people keep diaries and take pictures for that purpose. In this study, we kept things of personal significance in a time capsule for nine years and examined whether personal memories could be saved in a time capsule and how they might possibly change over time. We held a workshop in 2003 when participants contributed something that they possessed which had personal significance at that time of their life. They were interviewed to explain what kind of significance these possessions had for them, and these interview sessions were recorded. Nine years after

the initial workshop, the participants came together again. Before the time capsule was opened, they were asked to recall what they had put in the time capsule and to describe in what ways their possession in the time capsule had been significant to them. By comparing the contents of the participants' responses between 2003 and 2012, it was found that a great deal of the contents had changed from 2003 to 2012. Implications were discussed in regard to the significance of objects themselves and the narratives that go with the objects in preserving personal memories.

Published as: N. Shingaki, M. Kitabata, H. Matsuoka, T. Takada, A. Orito, Y. Kato, Y. Tsuzuki, and T. Owada, "Accuracy and Distortions of Personal Memories ("Omoide") Saved in a Nine-year Time Capsule," Cognitive Studies, Vol. 21, No. 1, pp. 15–28, 2014.

Papers Published in Technical Journals and Conference Proceedings

A Simple Method for Forming Compositionally Graded $In_xGa_{1-x}As_{1-y}Sb_y$ Base of Double-heterojunction Bipolar Transistors Modulating CBr_4-doping-precursor Flow in Metalorganic Chemical Vapor Deposition

T. Hoshi, N. Kashio, H. Sugiyama, H. Yokoyama, K. Kurishima, M. Ida, H. Matsuzaki, and H. Gotoh

Applied Physics Express, Vol. 7, No. 11, p. 114102, November 2014.

We studied a CBr_4-flow-modulation method as a way of simplifying the formation of a compositionally graded $In_xGa_{1-x}As_{1-y}Sb_y$ base of InP-based heterojunction bipolar transistors (HBTs) by metalorganic chemical vapor deposition. An investigation of C-doping in $In_xGa_{1-x}As_{1-y}Sb_y$ revealed that the In and Sb content decreases as the supply ratio of CBr_4 to group-III (R_C) increases. We fabricated 0.25-μm-emitter HBTs with a compositionally graded $In_xGa_{1-x}As_{1-y}Sb_y$ base formed by varying R_C at constant group-III, group-V, and V/III ratios. With this method, we obtained a higher current-gain cut-off frequency (504 GHz) and maximum-oscillation frequency (328 GHz) than those of uniform-$In_xGa_{1-x}As_{1-y}Sb_y$-base HBTs with the same base/collector thicknesses.

Channel Access Acquisition Mechanism Coupled with Cellular Network for Unlicensed Spectrum

R. Kudo, B. A. H. S. Abeysekara, Y. Takatori, T. Ichikawa, M. Mizoguchi, H. Yasuda, A. Yamada, and Y. Okumura

Proc. of VTC2015-Spring (2015 IEEE 81st Vehicular Technology Conference), Glasgow, UK, May 2015.

Interworking among heterogeneous wireless networks across licensed and unlicensed spectra has gained much attention as a way to handle the surge in mobile traffic. The wireless local area network (WLAN) is well known as the dominant wireless network application in the unlicensed spectrum. Unlicensed spectrum access should follow carrier sense multiple access/collision avoidance (CSMA/CA) in order to share wireless resources with existing WLAN nodes. However, the transmission throughput with CSMA/CA can be significantly degraded by the hidden terminal problem in environments where a huge amount of mobile traffic depletes the wireless resources. In this paper, we propose a channel access acquisition mechanism that uses licensed spectrum access. The proposed mechanism significantly reduces the impact of the hidden terminal problem by using both transmission and reception opportunities.

Experimental Verification of Highly-scalable OXC that Consists of Subsystem-modular Express Switch Part and Multicast-switch-based Add/drop Part Enabling Total Throughput of 314 Tbps

S. Takashina, H. Ishida, M. Niwa, Y. Mori, H. Hasegawa, K. Sato, and T. Watanabe

Optics Express, Vol. 23, No. 11, pp. 14796–14805, June 2015.

We propose cost-effective and scalable optical cross connect reconfigurable optical add/drop multiplexing (OXC/ROADM) that consists of a subsystem-modular express switch part and a transponder bank-based add/drop part. The effectiveness of the proposed architecture was verified via a hardware scale evaluation, network performance simulations, and transmission experiments. The architecture enables large throughput and offers significant hardware-scale reductions with marginal fiber-utilization penalty against the conventional architecture. Part of the OXC/ROADM designed to accommodate 35x35 express fiber ports and 2,800 transponders for add/drop was constructed. Its net throughput reaches 314 Tbps using 80 channels of 120-Gbps signals (where 30-Gbaud dual-polarization quadrature phase-shift-keying signals with 7% overhead are assumed).

Resource Allocation Scheme for Heterogeneous Traffic and Received Power in MU-MIMO-OFDMA Transmission

Y. Sakata, T. Murakami, Y. Takatori, M. Mizoguchi, and F. Maehara

IEICE Transactions on Communications, Vol. J98-B, No. 7, pp. 707–716, July 2015 (in Japanese).

This paper proposes a resource allocation scheme to cope with heterogeneous traffic and received power for multi-user

multiple-input multiple-output orthogonal frequency division multiple access (MU-MIMO-OFDMA). The feature of the proposed approach is to maximize the frame efficiency by allocating frequency and space resources under different user packet sizes and channel conditions. In allocating frequency and space resources, a large number of allocation patterns have to be considered. Hence, we reduce the number of patterns taking advantage of the fact that broadband channels in OFDMA provide almost the same performance regardless of the frequency band. Computer simulations showed the effectiveness of the proposed scheme in comparisons with MU-MIMO and OFDMA.

Excitation-inhibition Balances of Glx and GABA Predict Individual Differences in Perceptual Organization

H. Kondo, D. Pressnitzer, Y. Shimada, T. Kochiyama, and M. Kashino

Proc. of the 35th Annual Meeting of the Japan Neuroscience Society, Kobe, Japan, July 2015.

An essential function of perceptual systems is to structure the incoming flow of sensory inputs into a coherent scene. This is termed perceptual organization. Perceptual bistability provides us with clues to investigate neural mechanisms of perceptual organization because it produces dissociations between physical information and subjective experience. Here, we showed that the principle of excitation-inhibition balance is shared across auditory and visual bistability and independently implemented in the auditory and motion sensitive areas. We used magnetic resonance spectroscopy to noninvasively measure concentrations of glutamate-glutamine (Glx) and gamma-aminobutyric acid (GABA) in vivo. The time-series data of alternating percepts were obtained while participants listened to auditory streaming or observed visual plaids. Higher Glx concentrations induced shorter durations of alternating percepts, whereas higher GABA concentrations led to longer ones, regardless of sensory inputs. The two forms of neurotransmitter levels accounted for around 30% variance of percept durations for each modality. Our results suggest that the formation and selection of auditory and visual percepts depend on the opponency between excitatory glutamatergic and inhibitory GABAergic systems.

Luminance Profile Control Method Using Gradation Iris for Autostereoscopic 3D Displays

M. Date, T. Kawakami, M. Sasai, and H. Takada

Proc. of CLEO-PR 2015 (the 11th Conference on Lasers and Electro-Optics Pacific Rim), 26B3-6, Busan, Korea, August 2015.

A precise control method of angular luminance distribution of viewing zone using a filter with gradation in transmittance in an iris of a projector is proposed for autostereoscopic 3D (three-dimensional) display with smooth motion parallax.

Lagopus FPGA - A Reprogrammable Data Plane for High-performance Software SDN Switches

K. Yamazaki, Y. Nakajima, T. Hatano, and A. Miyazaki

Proc. of HOT CHIPS 27, pp. 10–19, Cupertino, USA, August 2015.

For cloud service providers and network service operators, software-defined networking (SDN) and network functions virtualization (NFV) are key technologies for automatic provisioning from an upper-management system and for enabling telecom operators to reduce CAPEX and OPEX. NTT has developed a high-performance SDN software switch called Lagopus, which has been released as open source software since July 2014. In this presentation, we reported on a Lagopus field-programmable gate array (FPGA) as a software-packet-processing-aware 40-Gbps FPGA NIC (network interface card) that was developed to fully utilize the multi-core central processing unit (CPU) power on the SDN/NFV platform with less than 10% x86 CPU power dissipation. We performed a live demonstration of the 40-Gbps wire-speed Lagopus FPGA at HOT CHIPS 27.

A 0.15-μm CMOS Baseband LSI Employing Sleep Mode with Clock-offset Compensation for M2M Wireless Sensor Networks

K. Suzuki, A. Yamagishi, and M. Harada

IEEJ Transactions on Electrical and Electronic Engineering, Vol. 10, No. 5, pp. 576–584, September 2015.

This paper describes wireless baseband large-scale integration (LSI) that contains a sleep management circuit. The sleep manager performs the sleep-clock offset compensation and enables a wireless terminal (WT) with a typical crystal oscillator to remain in sleep mode for a long period while maintaining synchronization with the access point. Lab experiments show that the sleep period reaches 512 s and that with intermittent operation the WT maintains synchronization with the access point for ten days. The LSI's average current consumption is as low as 11 μA for a 128-s sleep period. A wakeup detection circuit is also implemented in the LSI. This circuit performs paging control instead of a microprocessor unit (MPU), and this helps to reduce current consumption in the MPU and the flash ROM (read only memory). The single-chip baseband LSI is fabricated using 0.15-μm CMOS (complementary metal-oxide-semiconductor) technology. It is 4.6 mm × 4.2 mm in area and consumes 4.0 μA for sleep operation.

Viewpoint Image Generation for Head Tracking 3D Display Using Multi Camera and Approximate Depth Information

M. Date, H. Takada, and A. Kojima

Proc. of EuroDisplay 2015, p. 36, Ghent, Belgium, September 2015.

A simple and high image quality method for viewpoint image synthesis is proposed. Smooth motion parallax of wide depth range objects induced by viewpoint movement for left-and-right and front-and-back directions is achieved using multi camera images and approximate depth information. It is suitable for real-time 3D (three-dimensional) display applications.

Inscribed Fitting Method with Asymmetric Evaluation Function for Disaggregating Current Waveforms

F. Ishiyama, T. Watanabe, H. Inoue, and T. Ohyama

Proc. of IEEE 2015 ICCE-Berlin (the 5th IEEE International Conference on Consumer Electronics), pp. 19–23, Berlin, Germany, September 2015.

We propose a method for disaggregating the current waveforms of appliances from a current waveform on a power distribution board. We focus on the case where some appliances have current waveforms that vary continuously and non-proportionally. Our method is a variation of the least squares method with an asymmetric evaluation function. We calculate a standard waveform for each appliance, and

we inscribe the standard waveforms to the current waveform on the power distribution board. We apply our method to the aggregate current waveforms of three appliances and compare the results with the least squares method.

Top of Worlds: Estimating Time Complexity of Calculating Rank Order in Multi-dimensional Hierarchical Sets

T. Hata, H. Kawasaki, H. Kurasawa, H. Sato, M. Nakamura, and A. Tsutsui

Proc. of HASCA2015 (3rd International Workshop on Human Activity Sensing Corpus and its Application), pp. 1405–1412, Osaka, Japan, September 2015.

The increasing number of mobile devices such as smartphones has brought attention to participatory sensing, in which real-world data are collected via personal devices. To collect data via participatory sensing, it is important to motivate participants. Thus, we previously proposed Top of Worlds, a method for encouraging user participation by presenting their rank order. In this paper, we estimate its time complexity in order to understand how often we can present a rank order in the planning phase of services.

Large-scale Optical Switch with Simplified Sub-switch Connections for Datacenter Application

K. Ueda, Y. Mori, H. Hasegawa, K. Sato, and T. Watanabe

Proc. of Photonics in Switching 2015, pp. 366–368, Florence, Italy, September 2015.

We introduce an asymmetric-port-count delivery-and-coupling (DC) switch that can simplify fiber connection arrangement in a large-scale optical switch. A 24x4 DC switch is monolithically implemented with PLC (planar-lightwave-circuit) technologies, and its good performance is experimentally confirmed.

On the Computational Power of Constant-depth Exact Quantum Circuits

Y. Takahashi

Proc. of CFTM (Computability Theory and Foundations of Mathematics) 2015, Tokyo, Japan, September 2015.

We show that there exists a constant-depth polynomial-size quantum circuit for the quantum OR operation. We also show that, under a plausible assumption, there exists a classically hard problem that is solvable by a constant-depth quantum circuit with gates for the quantum Fourier transform.

Secret-key Distribution Based on Bounded Observability

J. Muramatsu, K. Yoshimura, P. Davis, A. Uchida, and T. Harayama

Proceedings of the IEEE, Vol. 103, No. 10, pp. 1762–1780, October 2015.

This paper reviews an approach to secret-key distribution based on the bounded observability (BO) model. First, the information-theoretic framework of secret-key agreement from a correlated random source is reviewed. Next, the BO model is introduced. In the context of this model, the BO condition is presented as a necessary and sufficient condition for the possibility of secret-key distribution. This condition describes limits on the information obtained by observation of a random object and models the practical difficulty of completely observing random physical phenomena. Finally, an implementation of secret-key distribution based on BO in an optical fiber system is described.
