Media Intelligence caring for people

Media intelligence technologies for knowledge processing and utilizing human speech and images.
Toshiba has long been working on research and development of media intelligence technologies, as featured in T-SOUL Vol.8.
After two years of further research and product development, we launched a cloud-type speech and image utilization service named "RECAIUS" in July 2015.
Following articles introduce "RECAIUS" along with its applied solutions and service practices.

Linking humans and things with cloud services

Accumulation of abundant technologies in speech, image and knowledge processing Leading the new era of media intelligence

Toshiba has been working on media intelligence since the 1960s. Many years of R&D allow us to provide services and products installed with high-performance engines which can process speech/language/image. By incorporating powerful machine learning represented by the deep learning that has been drawing attention in recent years, media intelligence is about to make further progress. Toshiba will take advantage of these technologies and contribute to creating a safe, secure and comfortable society towards the future that lies ahead in the era of IoT where humans, things and ICT interlink on a daily basis.

Hideo Umeki

Group Manager
Media Intelligence Product and Service Planning Group
Product and Service Marketing and Planning Department
Product and Service Marketing Division
Industrial ICT Solutions Company
Toshiba Corporation

Joined Toshiba Corporation in 1991. Engaged in R&D on neural networks and knowledge processing. After leading a R&D sector on speech recognition/ synthesis, translation, and dialogue, he is now responsible for promoting media intelligence project development.

Media intelligence is essential in the world of "Things × ICT × Humans"

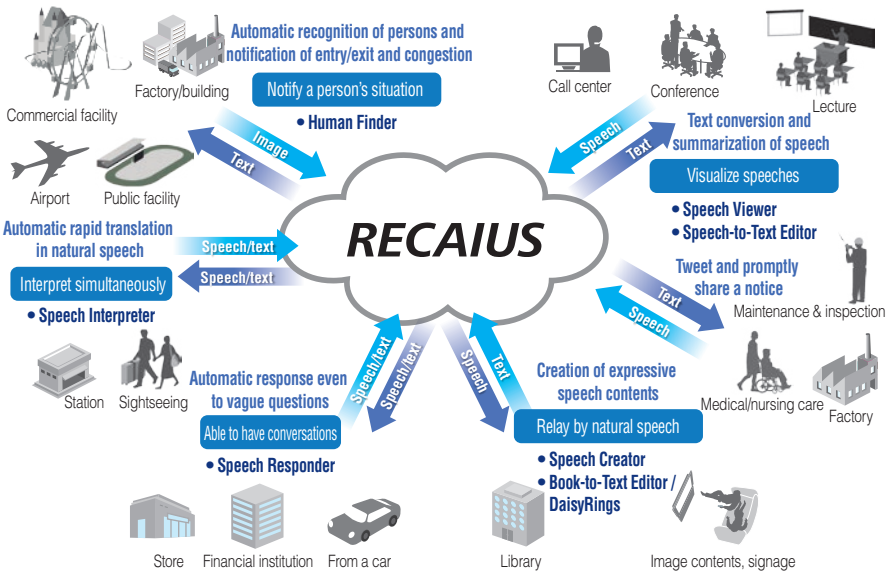
In the era of IoT[*1], new values are generated by interlinking things. And, we believe what lies ahead of IoT is "caring for people" ICT that supports people's lives and business activities safely, securely and comfortably through "Things × ICT × Humans" interlink.
There is the technology essential to understanding the intention and situation from human speech and behavior, and that is the media intelligence technology. Toshiba has worked on this technology since the 1960s, starting from the development of a postal sorting machine to automatically read handwritten zip codes.
There are roughly 3 major fields in media intelligence, namely, "Speech", "Image" and "Language and Knowledge". We have integrated these technology and started cloud services wholly named as "RECAIUS" which support an understanding of the intention or situation from speech or images, an easy communication and understanding among humans. We are aiming at creating a safe, secure and comfortable society with RECAIUS supplementing the capability to "watch, listen, and talk" possessed by humans.

Keep evolving the dictionary through learning and crowd-sourcing

At RECAIUS, speech and image data captured by various input devices are processed by a recognition engine on a cloud to perform intention understanding. In doing so, the engine refers to various dictionaries. The dictionaries play a key role in the processing system, and that is where Toshiba's strength lies. By combining the system to learn a massive amount of information collected from the Web etc. and collection and correction of information by crowd-sourcing[*2], the quality of media knowledge processing such as speech recognition is efficiently improved.
The main applicable sectors of RECAIUS are as shown below (Figure 1).
First, the sector to support field works, where RECAIUS can be utilized in sharing information and giving work instructions at the site of medical and nursing care or at factories and warehouses. Other than that, the sector to retrieve information from voice or image data, to respond to inquiries made by information navigation and to support communications by converting speech to subtitles or interpreting in voices. Additionally, the sector to support creating content such as advertisements and games by mainly utilizing speech synthesis technologies, and the sector to monitor the situation inside facilities by the figures of persons and images of faces.
There are 7 standard services to create such values. In addition to the 3 services that have already been provided, namely, "Speech-to-Text Editor", "Book-to-Text Editor/ DaisyRings" and "Speech Viewer", 4 more services are planned to be launched within this year.
The Speech-to-Text Editor supports work for writing down the content of speech, like the preparation of meeting minutes.

*1 IoT: Internet of Things
*2 Crowd-sourcing: To request many unspecified persons (crowd) to work on data processing, content creation, etc. and to obtain an outcome from a number of work results gathered.

Figure 1 Utilization examples of RECAIUS



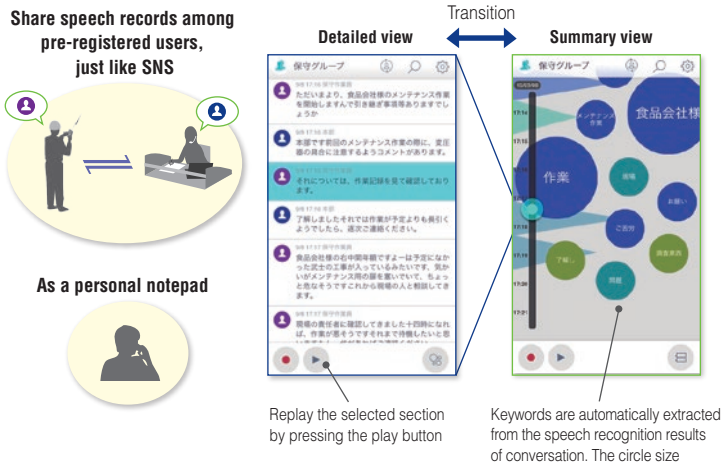
Toshiba aims at realizing a society where many people can live safely, securely and comfortably by creating new solutions for various social, life and business situations utilizing RECAIUS.

Through its unique functions such as showing of candidates for text by speech recognition, automatic replay of not yet written sections, and automatic speaker identification, the operation time can be reduced by about 30% on average.
The Book-to-Text Editor/ DaisyRings is a service to vocalize books using a speech synthesis technology, mainly targeting visually impaired persons.
Currently many text-to-speech services rely on a personal voice record by reading text, and it takes a long time to vocalize one book. Because of that, it is difficult to drastically increase the number of vocalized books. In the text-to-speech technique where a computer synthesizes speech from texts, the operation time becomes less than one third compared to recording human voices.
Viewing the Disability Discrimination Act that will be implemented in April 2016, similar needs may increase at public institutions and companies.
Before moving onto the Speech Viewer, let me explain the speech recognition technology that forms its foundation.
Toshiba has worked on R&D on speech recognition technologies for many years. By incorporating various cutting-edge technologies, lately, it has become possible to accurately recognize spoken words. For instance, when a speech in a lecture is recognized and converted to text in real time, a recognition rate of 85% or greater on average can be realized as long as specific terms are registered in the dictionary. Conventionally, realization of high-accuracy recognition required a massive amount of data for learning, and the cost of data gathering and learning has been the hindrance to introducing such a system. To that end, at RECAIUS, a technology which enables making a highly accurate dictionary from a small amount of data for learning was established by preparing a large-scale text database. Additionally, there is a system to semi-automatically register trending words into a dictionary using crowd-sourcing. This is also utilized for updating the dictionary of the speech recognition engine in Toshiba's 4K TV REGZA "Z10X" Series' program search and other functions.

Swift access to information by multi-scale summarization

Speech Viewer is a service to visualize speech data which were created by fusing this speech recognition technology and information summarization technology.
To date, for business use, speech recognition has been utilized in text conversion and analysis of many hours of telephone conversations mainly at call centers (see Page 9). For general use, it has been limited to a supplementary input system of text input used in smartphones. By improving its accuracy, customizability and operation efficiency, speech recognition may become possible in the future to assist human communications utilizing recording and recognition of many hours of speech at a wider variety of business and life occasions including lecture sessions, conferences and SNS.
However, a simple text conversion of many hours of speech data results in a long sequence of non-structured texts, which is very difficult to handle. The key point here is the information summarizing technology for facilitating the human perception of abundant information. That is also used for analysis of the content of conversations at call centers. For example, information shown on a scalable map changes depending on its scale. When zoomed in, detailed information such as building names are shown, while when zoomed out, only highly important information such as the municipality name will remain displayed. Such a concept of multi-scale information summarization is also very effective in understanding abundant speech data. Our Speech Viewer is created by fusing speech recognition and multi-scale summarization, where the temporal axis of speech data corresponds to the spatial axis of a scalable map.

Figure 2 Conceptual illustration on utilization of RECAIUS Speech Viewer and example screen images



Speech Viewer recognizes and visualizes (converts to text) speeches, and assists to grasping the entire conversation by visually displaying the keywords.

In this Speech Viewer, even if many hours of speech data are accumulated, the user can get an overview of the entire conversation on a screen where keywords are extracted and visualized automatically and search for speeches by selecting the relevant keyword. Additionally, the system allows for checking the content by listening to the actual speech along with the text created by speech recognition (Figure 2).
Through these functions, the Speech Viewer is expected to provide new ways of utilization in various business situations such as casually exchanging ideas and notices among employees with many field work hours and sharing the gist of meetings in the form of speeches instead of the written minutes, in addition to grasping the content of speeches over many hours in, for example, presentations and lecture sessions.

Highly expressive speech synthesis and high performance speech/human recognition

Services that will be provided in the future are “Speech Creator”, “Speech Interpreter”, “Speech Responder”, and “Human Finder”.
Speech Creator is a speech synthesis service with diverse emotional expressions. It currently supports 11 languages including Japanese, English, Chinese and Korean. The speech synthesis engine of the Speech Creator is also utilized in the voice navigation system of the popular smartphone App “Yahoo! Car Navi”.
The excellent points of Toshiba’s speech synthesis technology are its high voice quality and rich expressiveness. It takes only one hour in processing a 30-min voice recording data for the system to be able to learn the speaker’s characteristics in speech and vocalize any text in a similar tone and expression.
The technology of speech interpretation is also advancing. Conventionally, in speech interpretation, it was difficult for users to have smooth conversations because users had to input each sentence one by one. Our Speech Interpreter can recognize, translate and display a natural speech in a conversation simultaneously and continuously without any break, so the quality of communication has been drastically improved (see Page 6).
Speech Responder refers to a function to automatically make correct responses by understanding the intention of the speaker. For example, when it recognizes a speech saying “the address was changed” or “I moved to a new place” at a contact center, it will show the procedures of notifying an address change automatically. In general, computers are not very good at handling diverse expressions or vague expressions. However, by efficiently collecting various expressions through crowd-sourcing and by making the computer learn intention-understanding models, the linguistic level of Speech Responder was substantially increased.

Leaving what edges can do to edges Solving the issues of privacy and network

The last service on the list is Human Finder. The technologies used here are facial and human image recognition (see Page 12). This technology judges the intention or situation related to human activities by combining a technology to identify persons and a technology to detect the figure or movement of persons reflected to a camera.
One of examples of its use is to install these cameras at a shopping mall to visualize the movement and density of customers and to take action based on the information, such as to move display racks to more populated areas. It may also be possible to recognize VIP visiting an event venue and to notify the event manager of it.
However, accumulation of image data including people’s faces on a cloud requires extra attention in terms of privacy. There also is an issue of overloading the network since image data are generally large in file size.
Therefore, we will leave what edges (device side) can do to the edges. By sending the only processed data to the cloud after simple image processing on edges, the issues of privacy and network can be solved. The image recognition processor that

plays an important role there is "Visconti2" developed by Toshiba. A camera with Visconti2[*3] can be applied to analysis of customer movement lines and grasping the situation of congestion.
In this way, RECAIUS is able to handle various forms of data such as speech, image and text in an integrated manner. For future examples, robots and humans naturally communicate to each other, complicated situations are explained with easy-to-understand speech and accumulations of specific knowledge answer various questions automatically, etc. RECAIUS will further evolve into new services in the future where interface technologies and knowledge processing are fully fused and integrated.

Pursuing "customer friendliness" by ease of customization

Osamu Hori
Director
Corporate Research & Development Center
Toshiba Corporation

Toshiba's R&D Center is in charge of R&D on technologies that form the basis for all the business activities developed by Toshiba's companies and group enterprises. While the R&D departments of the companies and group enterprises have short-term achievement targets, we work on mid- to long-term R&D activities.
For the technologies of media intelligence, we have worked on R&D since the 1960s with a long-term point of view. The application of media intelligence has been limited to some selected sectors to date. However, in the last several years, there was a major breakthrough in the technology. That is the cutting-edge machine learning method called "deep learning" which achieved this breakthrough. By utilizing the deep learning along with the technologies and know-how accumulated over years, we realized the world's top-class accuracy and performance in various media intelligence fields including speech recognition, speech synthesis, and image recognition.

Realizing the accuracy required for business use by easy customization

The cloud-based service "RECAIUS" delivered by the Industrial ICT Solutions Company is filled with the outcome of this technological breakthrough. We also took part in the development of RECAIUS. For instance, assuming that the initial speech recognition engine was embedded into devices, we have focused on a small compact design. However, RECAIUS is a cloud-based service, and many users use it for various purposes at the same time. For that reason, to meet the relevant business divisions needs to use a high performance speech recognition engine, such as improved vocabulary of dictionary referenced by the engine and higher processing performance, we finally realized the service.
While a majority of global companies providing services using speech recognition technology aim at a generic speech recognition engine, Toshiba's directionality is different from that. We aim at utilizing a speech recognition technology which supports spoken languages and industry-specific terms at business situations where higher accuracy is required, such as at call centers. Easy customization is the key of feature.
Toshiba's media intelligence technologies have strengths in the basic performance as well as the dictionary that can be customized at low cost in a short period of time.
It is our desire to continue working on the development of technologies that serve as the strengths of Toshiba and research that supports its future business while being always aware of customer needs and the market environment, in collaboration with the Industrial ICT Solutions Company, business divisions of the Toshiba Group, and other R&D divisions.

*3 Visconti2-installed cameras are in the stage of sample production as of October 2015.

"Omotenashi" by "Caring for People" services

Total inbound services for welcoming ("Omotenashi") overseas visitors to Japan

Before the upcoming Tokyo Summer Olympic Games in 2020, expanding and improving services for overseas visitors to Japan, so-called inbound services have been drawing much attention in Japan. One of serious problems there is how to communicate with foreigners, and the speech interpretation services are expected as a solution to overcome the language barrier. Based on advanced language processing technologies accumulated over many years, Toshiba is now working on the development of various inbound services. What we are aiming for is creating new "Things plus Experience" that support both the overseas visitors and Japanese hosts.

Atushi Sakuma

Chief Specialist
Technology Group
Media Intelligence Business Development Department
IoT & Media Intelligence Business Creation Division
Industrial ICT Solutions Company
Toshiba Corporation

Joined Toshiba Corporation in 1992. In charge of multiple corporate customers as Technology Chief Specialist on system solutions designed for distributions and financial business. Currently engaged in planning and development of new business centering on distributions and financial sectors.

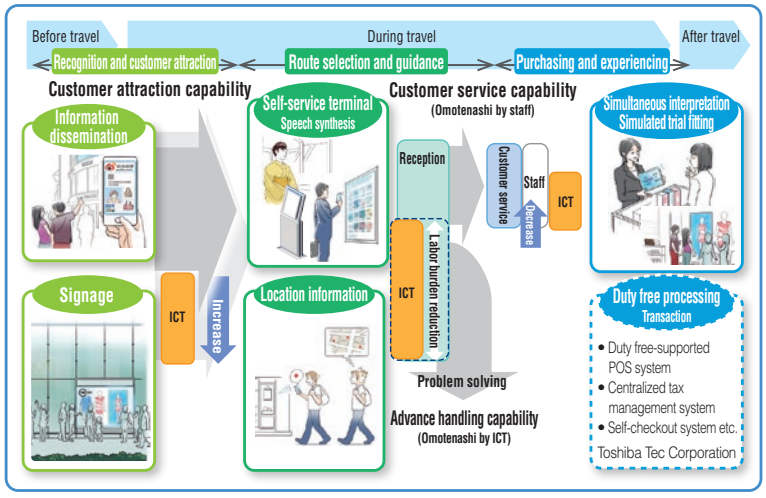
Aiming at creating "Things plus Experience" for inbound services

The number of overseas visitors to Japan in 2014 reached 13.41 million, which was 29.4% more than that in the previous year[*1]. As the Japanese government sets the goal of increasing the number of overseas visitors to 20 million, the number of visitors is expected to increase further in the future. The Tokyo Summer Olympic Games will be held in 2020. It could be said that expanding and improving the services for overseas visitors to Japan is a major task for the entire Japanese society. We are thinking to play an important part with new services in which our original technologies are utilized.

"Let the visitors get to know local Japanese communities and feel the values that are unique to Japan." For these reasons, we currently are aiming at creating inbound services for providing overseas visitors with "fun", "surprise" and "excitement", not just "convenience". While experiences such as sightseeing, shopping and dining are important, we want to realize "Omotenashi" (Japanese-style hearty hospitality) through extra services. That goes in line with the directionality of creating "Things plus Experience" held up by Toshiba Corporation.

Inbound services consist of a wide variety of elements. The summary structure of them is that various services can be prepared on the inbound service infrastructure by linking devices such as sensors, smartphones and digital signage through the network. "RECAIUS" realizes speech interpretation and intelligent dialogue and plays the major role as a cloud service in the service infrastructure. By adding other elements such as location information, authentication and settlement, the infrastructure supports a broad range of inbound services comprehensively. Here, technologies on speech interpretation and location information become exceptionally important. The main customers of our inbound services are those at the hosting side, such as real-

Figure 1 Omotenashi by total inbound services.



For the three phases "recognition and customer attraction", "route selection and guidance" and "purchasing and experiencing", the total inbound services provide support for both the overseas visitors and Japanese hosts.

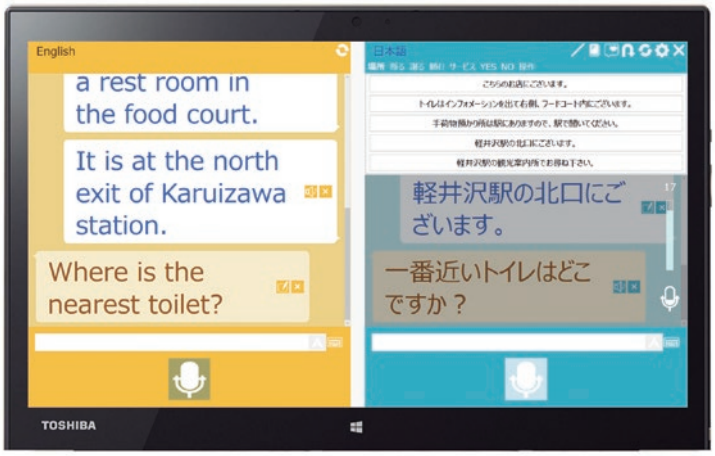
*1 "Foreign Visitors to Japan by Nationality and Month (for 2003-2015)" by Japan National Tourism Organization (JNTO)

estate developers, retailers, transport facilities and hotels.
Services at such commercial facilities are roughly divided into 3 phases. They are, “recognition and customer attraction” for improving the customer attraction of stores, “route selection and guidance” for reducing the anxiety or stress of foreigners at facilities and areas, and “purchasing and experiencing” for supporting and improving the capability of staff who deal with foreigners (Figure 1).

Speech Interpreter to support communication between store staff and customer

In the phase of recognition and customer attraction, it is important to provide proper information at the right time and place. For example, local promotion to attract foreigners to visit Japan in the stage of planning, information delivery to smartphones, and showing information of potential interest on digital signage by detecting the gender and age of the person standing in front of it using a camera.
In the phase of route selection and guidance, we can show you a good example of information provision which uses self-service terminals. When an overseas visitor asks a question to a signage in a facility, the signage will give the answer in the mother tongue of the visitor. Of course text information can be displayed, but in some cases synthesized speech may be more suitable. It is also possible to guide the person to the right direction with smartphone by recognizing the location information. Through the ICT that supports route selection and guidance, the burden on the staff of facility information and help desk will be alleviated. Looking from the viewpoint of stores and facilities, enhancing the “advance handling capability” allows for improving the efficiency of facility operation and the overall quality of “Omotenashi”.
In the phase of purchasing and experiencing, inbound services provide support for smooth customer services. To date, the services have been introduced to and verified at various places such as local shopping malls and department stores in Tokyo. Due to the diversifying nationality and language of visitors in late years, shopping malls and department stores started to struggle with communication issues. To resolve the issue, the Speech Interpreter service we developed was provided (Figure 2). Using tablet terminals, conversations between store staff and a visitor are translated in real time.
The Japanese language spoken by the store staff and English language spoken by the visitor were instantly translated and shown on the tablet at hand. We could see some outstanding outcomes of the features the system has. For instance, the store staff can answer only in Japanese to inquiries made by the visitor who speaks English, and these conversations are displayed on the tablet in both languages for their peace of mind, in addition, they can have smooth conversation by using a dictionary containing words and phrases often used in conversations at the store. This service is currently available in Japanese, English and Chinese. In addition to such a stationary-type speech interpretation service using tablet terminals, we are giving considerations to providing a mobile service using a smartphone app and an automatic service by unattended terminals.

Figure 2 Example screen of the Speech Interpreter service.



A conversation between a visitor and store staff is translated in real time and displayed on the screen (English on the left and Japanese on the right in this example).

We’ll try to improve inbound services further by revealing hidden issues of Omotenashi for overseas visitors to Japan through such trial services.

Combining highly accurate speech interpretation system and support by interpreter

Our Speech Interpreter service has its strength in the dictionary. The dictionary can be customized to support the linguistic characteristics of a business or store, local trends, as well as added with technical terms, industry terms, or company-specific terminologies. Our Speech Interpreter service can be utilized not only for general consumers like at retailers but also for highly specialized business.

However, though the accuracy of speech interpretation has been improved, it is evident that there currently is a limitation in improving the accuracy only by computer technologies such as machine learning. To that end, we are considering supplementary services utilizing humans. What we are thinking is, for instance, when there are occasions like "I can't make myself understood" or "I can't understand the nuances" in the exchange with the Speech Interpreter service, the user will be directed to a center staffed with interpreters to receive support. With this, the original task of "better communication with foreigners" will be realized, and the host side will be able to deal with foreigners with less staff.

Activities for improving the accuracy of the Speech Interpreter service never cease. Every time the service is used, individual expressions and points of improvement are fed back to the system, and the dictionary improves itself through the process of machine learning.

Accumulation of such a big data will allow us to utilize the system in multiple fields in the future. For example, by systematically summarizing troubles and difficulties faced by foreign visitors, it can be utilized for developing new products and services specifically designed for foreigners. Linking to location information, we may be able to propose to improve the attraction of a region by analyzing conversations made at various locations in the region.

Although the number of foreigners have been increasing, the current inbound services available in Japan are not enough. We are thinking to launch new services in the future through collaborations with companies that have various customers and technologies.

Business innovation utilizing speech recognition

Improving CRM by visualizing voice of customer Quickly understanding market changes and customer needs

Many companies are enhancing their customer relationship recently. Among customer relationship, contact centers play a substantially important role. Various companies are reviewing the existing operation and ICT system, aiming at establishing next-generation contact centers. While there are various ways of approaching to achieve that, utilization of speech recognition is currently drawing attention as one of ways to do this. Fusing the speech recognition technology developed over years and technologies for summarizing and analyzing speech converted to text, Toshiba brings new developments to contact center solutions.

Ayumu Shimizu
Ph.D. (in Engineering)
Specialist
Customer Experience Management Products and Service Engineer
Product and Service Marketing Division
Industrial ICT Solutions Company
Toshiba Corporation

Joined Toshiba Corporation in 2002. After engaging in R&D on network security and green IT, he took charge of CRM solutions from 2013, promoting introduction of new technologies etc.

Speech recognition that leads CRM systems

Contact centers receive various inquiries, requests, opinions and complaints of customers. While responding to them, the operators report the key points to the company through the system. Storing the answering record is an important job in order to share the information in company. However, since it is often required not only to improve the quality of it but also to shorten the time of the inputting work, entry omissions sometimes occur actually. If information that indicates defects of a product is left without being revealed at the contact center, there is a risk of delay in finding the issue and the situation becoming worse.
The reason why the growing number of companies introduce speech recognition is that they have needs to minimize such a risk or enhance the compliance. For example, if a phone conversation is converted to text by speech recognition, an inappropriate response can be easily found and corrective measures can be taken accordingly. It may also be useful in giving guidance or training for operators and reviewing their work.
If we call it "defensive" measures, "offensive" measures would be planning and developing new products or services by gathering VOC[*1]. Converting VOCs gathered at a contact center to text by using speech recognition and sharing them among relevant departments in the company may lead to obtaining customer insights and creating new ideas earlier than before[*2].
Toshiba has long been working on the field of media intelligence and has an advantage in speech recognition technology. Speech recognition technology for call center operations has reached a level where it is ready for practical only after some minor tuning. But in the practical use of this speech recognition, the simple text conversion of speeches is not enough for the work at contact centers, and the combination with peripheral technologies are important.

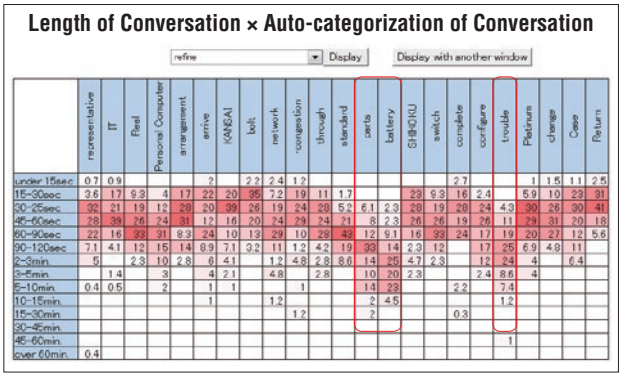
Text mining for promoting utilization of speech recognition technology

When all voices in a conversation are fully converted to text, it becomes difficult to understand the context. Human's speech often contains some meaningless interjections like "well", "umm", responses without influence on the context and unnecessary words. If the entire speech containing these is converted to text, it usually becomes very difficult to read.
To that end, there is a necessity for a technology to summarize the conversation converted to text, to improve the usefulness of speech recognition data. Toshiba has also been working on the development of such technologies in stages. By applying the current summarization function to conversations at a contact center, we could reduce the number of words substantially. It is also possible to adjust the word reduction level as necessary. We are currently working on enhancing the summarization function using AI, and it is anticipated that the number of words can be reduced to less than half of the current state (less than 30% of whole speech converted to text) in the near future while maintaining the context of conversation. Through this, it will be easier to understand the context, and use of speech recognition results will be promoted.

*1 VOC: Voice Of Customer
*2 Described in detail in Figure 1 on Page 6 of T-SOUL Vol.14

Furthermore, by analyzing the text groups, it may become possible to obtain useful VOC and unnoticed findings. What plays a key role there is the text mining technology. Toshiba has been working on unique activities in this field as well.
For example, when conversations at a contact center are processed by speech recognition and automatic keyword extraction and sorting, it will become possible to analyze the content of inquiries. It is not necessary to set keywords in advance, and it may derive new categorizations that we have never thought of.
Figure 1 shows the analysis results at a contact center of a maintenance service company. This result clearly shows some trends, like conversations containing keywords "component", "battery" and "fault" tend to become longer.
Through everyday observations, the managers of the work site might have felt something like "Operators are having hard time when the inquiry is about a battery". There may be some managers thinking about the necessity for improving the manual about batteries. However, for such a thought depending on qualitative evaluation, managers may be uncertain and hesitant about taking a step to make improvements. In such occasions, text mining should be able to assist in making a proper decision by presenting quantitative indicators through analysis.
Additionally, viewing transition of keywords in the speech text created by speech recognition alone may be useful in understanding the market trend. By accurately examining the number of telephone conversations containing trended keywords at the contact center of a company, that company should be able to learn whether its products match the latest interests of the market. Furthermore, by performing automatic clustering periodically, it is possible to read the change in market from the changes in clustered words.

Figure 1 Example analysis results of conversations at a contact center



Telephone conversations undergo speech recognition, where keywords are automatically extracted and categorized. Clearly showing a trend that conversations containing "component", "battery" and "fault" tend to become longer.

Like these examples, one of the strengths of Toshiba's contact center solutions is the ability to provide thorough services from speech recognition to analysis of results by text mining.

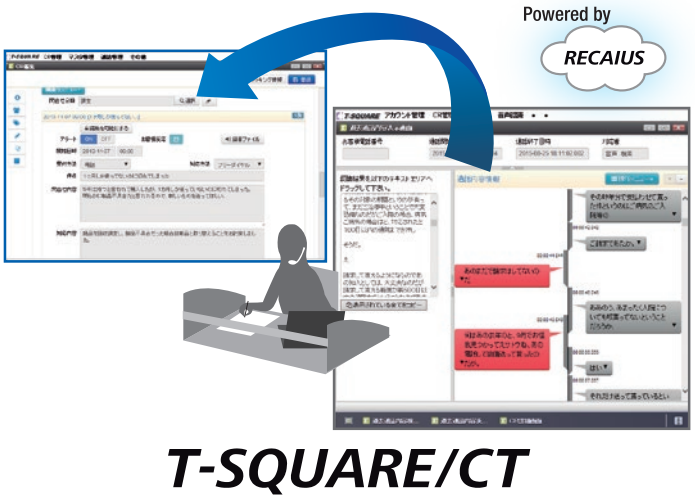
Real-time speech recognition improving the quality and efficiency of worksite operation

The method of speech recognition described above is rather suited for batch processing. By introducing real-time processing, the use of the speech recognition is expected to expand much further.
At many contact centers, supervisors perform monitoring of multiple operators. These supervisors check the content of conversations the operators are making, and provide support as necessary. However, since it is extremely difficult for one supervisor to listen to multiple telephone conversations at the same time, and it may not always be the case where the operator who is in dire need of support can receive assistance by the supervisor.
If real-time speech recognition is introduced, such a situation can be greatly improved. By recognizing speech and immediately displaying the text on the screen of a supervisor, it becomes much easier for the supervisor to monitor multiple operators. Additionally, it is possible to specify "risky words" and emphasize in red when such words are spoken, which facilitates improving the quality and efficiency of monitoring.
Meanwhile, in a conversation with a customer, an operator has to think "What kind of answer is most suited" and "What would be the appropriate suggestion" while trying to understand the intention of the customer. Every time such necessity arises, the operator seeks for an ideal answer by referencing various materials (paper documents or electronic data) available in the company. It takes a certain amount of time for the operator to find the needed material. If a system automatically recommends multiple materials that are relevant to the content of conversation, the operator can browse related information by simply selecting the material among the recommended materials, and it will improve the quality and efficiency of response.
The technology of speech recognition is steadily and constantly developing. Besides improving the recognition rate, when considering its utilization, collaborated use with other systems such as relevant material automatic recommendation engine is also an important task. We provide "T-SQUARE speech recognition option" where the "T-SQUARE/CT" (combination of

function units related to contact center operations among the CRM solutions "T-SQUAREx") and the speech recognition and peripheral functions described above are combined[*3]. Since they run on the same platform, speech recognition and the CRM system can seamlessly collaborate with each other (Figure 2).

Using this speech recognition-incorporated CRM system, we will support establishing next-generation contact centers. It is our desire to lead in creating contact centers of the new era by further advancing various technologies directly related to speech recognition and peripheral technologies to be collaborated.

Figure 2 Collaboration of speech recognition and CRM system



Toshiba's speech recognition and CRM system seamlessly collaborate on the same platform.

*3 The overview of T-SQUAREx is shown in Figure 1 on Page 4 of T-SOUL Vol. 14

Safe and comfortable society brought by image data analysis

Architecture of image data processing for reducing the load on network or cloud

Technologies of image data processing have been utilized in video production for TV programs and movies, improving the performance of digital cameras and in analysis of images taken by surveillance cameras, and have been developed in the process. Toshiba has continuously worked on R&D on image data processing for its various products such as TV systems and business, and the scope of its application has been expanding ever wider. What we are specifically focusing on in recent years is analysis of image data of "humans". Based on its facial and human recognition technology, Toshiba will provide various solutions that understand the human intentions and surrounding situations. Some of these latest technologies and Toshiba's unique activities are featured in this article.

Yasuhiro Taniguchi
Ph.D. (in Engineering)
Chief Specialist
Media Intelligence Product and Service Planning Group
Product and Service Marketing and Planning Department
Product and Service Marketing Division
Industrial ICT Solutions Company
Toshiba Corporation

Joined Toshiba Corporation in 1995. After being previously engaged in R&D on robot vision, on-board image recognition, image recognition LSI, etc., he currently promotes planning and development of new products where image recognition technologies are applied.

Comprehensive media intelligence with both intention understanding and situation understanding

Our eyes and ears play a central role in understanding other people. For instance, when we hear a person saying "I'm hungry", we understand the person "wants to eat" (intention understanding). However, in many cases it is difficult to make a proper response only by listening to someone speaking. The response should be different depending on the situation if the speaker is alone, a couple or an entire family. Human eyes can instantly grasp such a situation (situation understanding).

By combining both intention understanding and situation understanding, we are able to have better communication. Based on such a concept, Toshiba has been making efforts in the development of media intelligence where speech data processing, image data processing and knowledge processing are all fused and integrated.

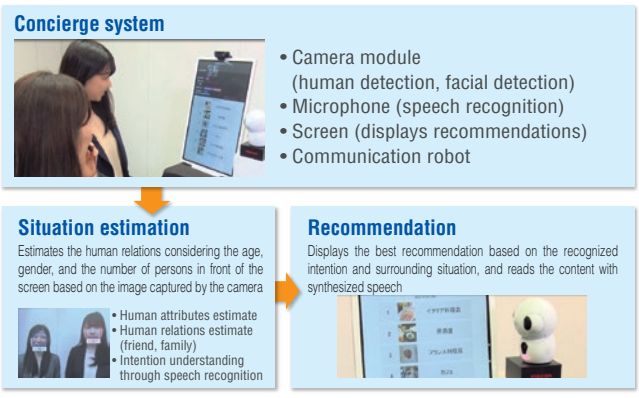
Figure 1 shows a simple example of comprehensive media intelligence solutions. When a person talks "I want to eat something" to a microphone, a camera constituting the system recognizes the situation, and the system recommends restaurants that considered to be the most suitable for the person/s. There, the information displayed for 2 women may be different from that for a couple consisting of a man and a woman.

When it comes to the image recognition, its application was almost always for the security sector in the past, like entry/exit control and crime prevention by face authentication. Recently, in addition to the security purposes, there is a strong wave in utilizing image recognition in much wider sectors.

For example, solutions for crime prevention and marketing are realized by connecting devices such as IP cameras and cloud services. Or, it is utilized in preventing collisions and improving the safety of automobiles by using on-board image sensors. There may also be a way to use image recognition to monitor deterioration of social infrastructures such as bridges by installing a camera to a drone. Among these various uses, what Toshiba is specifically focusing on are two technologies related to "humans".

One of them is facial recognition technology. This technology includes face detection for extracting the facial area from an image, face collation for identifying the person, and face attribute detection for distinguishing, for instance, the age, gender, and the degree of smiling from

Figure 1 Concierge service with integrated speech technologies



Providing the user with the most suitable information based on the results of intention understanding by speech recognition and situation understanding by image recognition.

the face. To recognize a face, first, the face of person is detected from an image, and the characteristic points such as the eyes and nose contained there are detected. Then, face collation is carried out after making corrections for the face direction and lighting. For the face direction, the system can identify the face by correcting up to 30 degrees horizontally and up to 15 degrees vertically.
Another is human recognition technology. This technology includes human detection for detecting the upper half or the entire body of a human from an image, human tracking for tracking one person in a screen, and human collation for collating a person based on the face, clothing, etc. Human tracking is a technology to track a person in a screen, but by combining with human collation, it can be used to track a person across multiple cameras. Wide-area human tracking may be applied to the fields of crime prevention and marketing.

Findings from signage audience rating survey Attracting attention to advertisement requires contrivances

We would like to talk about signage audience rating survey as an example of implementing our solutions that were realized by combining the technologies described above.
This survey was conducted by installing a camera next to a digital signage screen and quantitatively measuring how many people paid attention to the advertisement. For each audience, the facial attributes such as the age and gender were detected, and the level of attention to the contents on the screen was measured. The level of attention was divided into several grades such as "Passed through", "Glanced briefly", "Looked for more than 2 seconds while walking", and "Stopped walking and looked for more than 2 seconds" and scored. By summarizing and analyzing the results by attribute and score, it can be used for, for example, creating an advertisement with better effects or switching the content in accordance with the targeted audience.
Toshiba also provides displays. The mirror signage system manufactured for the demonstration experiment looks like a mirror when the power is off. When the power is turned on, contents are displayed and the installed camera remains invisible from the outside. Although careful consideration to privacy is required, this system is ideal for capturing the natural behavior of viewers.
We could find an interesting result from a demonstration service conducted at the Nihonbashi Mitsukoshi Honten in April and May 2015. It was a survey on the reaction of audiences conducted by combining the humanoid communication robot "Aiko Chihira" developed by Toshiba and signage. Aiko-san was there to give explanations on the content displayed on the signage. The first 3 days, we displayed the contents on the signage without Aiko-san, passengers barely reacted. The number of audiences ranged from several dozens to a little over 100 per day. The next 2 days, Aiko gave explanations on the content. The number of viewers reached over 2000 in the first day, and neared 5500 in the second day. It seems one contrivance or another is required to attract people's attention to signage.

Processing images by on-site cameras and devices Reducing load on network to less than 1/100

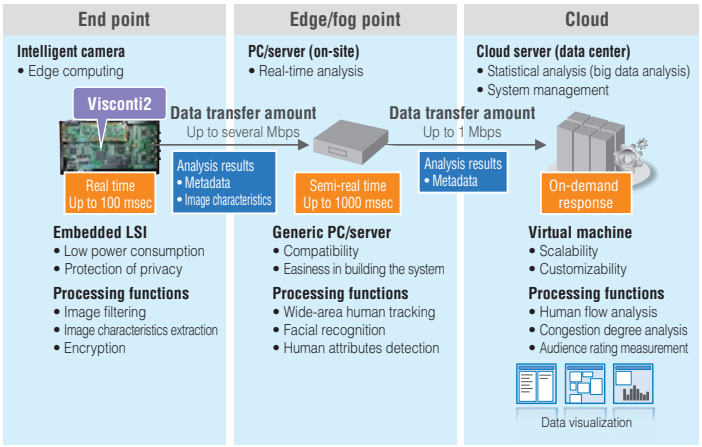
Next, we would like to talk about the experiment of a large-scale measuring system of congestion rate that was conducted at exhibitions and other venues. The mechanism of the system is to visualize congested places in a form of a heat map using multiple cameras by combining the technologies of human detection and tracking and the map data of the area.
By measuring the congestion rate, it becomes possible to take a wide variety of measures. For example, the flow of people is directed to relatively uncrowded booths. Places that tend to have stagnant flow may require some countermeasures. It also becomes easy to make improvements such as ideal allocation of staff and changing the direction of booth signboard considering the direction of traffic.
Through these demonstration experiments, it was discovered that the protection of privacy related to facial images and the data transfer amount when the number of cameras is increased may pose problems in practical application of image recognition systems in the future.
Surveillance cameras at public spaces, stations and buildings are already recognized as systems necessary for society and rules have been established on their operation and data handling. Similarly, also for image recognition systems, rules on their operation may be established in the future in order to take the balance between their usefulness and protection of privacy.
Regarding the matter of the data transfer amount, we believe the amount of data transferred to the cloud via the network can be reduced by processing images at the on-site devices and servers installed at the edges. Practically speaking, transferring each individual video clip with large file size to the cloud is not very realistic. When capturing an HD image, one camera

consumes a network bandwidth of about 10 Mbps. If 100 cameras are installed in a facility, that adds up to 1 Gbps. Such a heavy traffic of data puts pressure not only on the network but also on the computing power of the cloud.

What plays an important role in solving such issues is the image recognition processor "Visconti2" developed by Toshiba. When Visconti2-installed cameras[*1] are placed at the site and a device with a PC-like computing power is placed at the location of a hub for dealing with several dozen cameras, the amount of data transferred to the cloud will be less than 1/100 (Figure 2). If images are processed on site like this and only the metadata of image like "20s, Male" are accumulated on the cloud, the concern for privacy may also be eliminated. Our ambition is to present a rational image data processing architecture for saving the resources of the network and cloud as well as to further pursue the potential of media intelligence.

Although I mainly talked about image recognition technologies here, our media intelligence services aim at realizing the new IoT world that links the real world and the Internet world by integrating speech technologies, image technologies and robot technologies Toshiba has been diligently working on.

Figure 2 Configuration of cloud collaboration system in image recognition service



Both the reduction of the amount of data transmission and the protection of privacy are realized by edge-oriented image recognition service utilizing high performance image recognition LSI.

Published by: **Toshiba Corporation**

URL: http://www.toshiba.co.jp/cl/en/

E-mail: INS-info@ml.toshiba.co.jp

The feature articles herein are translations from the Japanese edition of *T-SOUL*, Volume 16, a Toshiba journal.

*1 Visconti2-installed cameras are in the stage of sample production as of October 2015.