Studying Multimodal Interactions: Understanding Dialogue Mechanisms through Combined Observation of Speech, Language, and Body Movement

Ryo Ishii
Distinguished Researcher, NTT Human Informatics Laboratories

Overview

Digital Twin Computing is one of the three key fields of technology in the IOWN (Innovative Optical and Wireless Network) vision. In order to build a digital-world representation of a person that includes not only their external aspects, but also their internal aspects, such as consciousness and thought, it is essential to understand and model the mechanisms of human communication. In this article, we speak to Distinguished Researcher Ryo Ishii, who works with multimodal information including speech, language, and body movement with the aim of understanding the mechanisms of human communication, for example in how a person conveys their mental state.

Keywords: human communication, multimodal interaction, dialogue system

Communication is born of the interaction between multiple modalities such as speech and gestures

—What kind of research field is multimodal interaction?

When people talk with others, they communicate using more than just their voice and language. They use multiple modalities, such as gaze, facial expression, and gestures, communicating what we call "multimodal information." These modalities are used in combination, and the modalities influence each other as people use them to communicate information. These interactions using multimodal information are called "multimodal interactions." What's particularly important about these interactions is that the transmitted multimodal information is dealt with as a whole. For example, during a conversation someone might say, "Don't joke about that." Looking at the written words alone, you can't tell whether the speaker is angry or joking. But when you listen to the

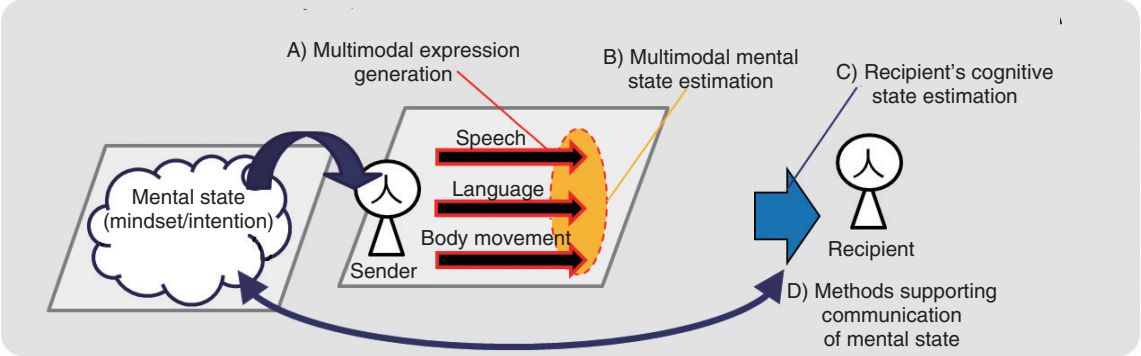

Fig. 1. The field of multimodal interaction.

tone of their voice and look at their facial expression, you know that if they are speaking softly with a big smile, they're probably not serious. One major research topic is comprehensively processing this multimodal information to understand the intentions and messages that people convey, and understanding and modeling the mechanisms for how this information is transmitted between people. Another major research topic in multimodal interactions is using this understanding and these models to support communication between people and enable smooth communication between people and dialogue systems.

In this research area, I am carrying out a multifaceted study of mechanisms for expressing and recognizing multimodal information (**Fig. 1**). This involves A) "multimodal expression generation," which is researching and modeling the mechanisms for generating speech, language, and physical motion to transmit the current mental state; B) "multimodal mental state estimation," which is estimating the mental state of the sender through their speech, language, and body movement; and C) "recipient's cognitive state estimation," which is looking at how the recipient interprets the sender's mental state. In addition, I am carrying out practical research into D) "methods supporting communication of mental state," which involves recognizing discrepancies between the sender's state of mind and their state of mind as perceived by the recipient, and enabling the sender's state of mind to be correctly conveyed to the recipient.

—What are you researching specifically?

As an example, one topic I worked on at NTT Communication Science Laboratories was predicting the next speaker and timing of speech. In the future, dialogue systems will need to quickly predict who should talk next, enabling them to speak at the appropriate time during conversations with people. So, we built a model that predicts who will speak next and when, based on subtle behavior like people's eye movements, head movements, and breaths taken before speaking. Since there were no existing studies that measured people's respiration during a conversation and analyzed it in relation to speech behavior, our paper presenting this research was given an Outstanding Paper Award at the ACM International Conference on Multimodal Interaction, the leading conference in this research field.

In addition, we work closely with internal and external researchers to conduct a variety of research related to multimodal interactions. This includes work on the "MoPaCo Window Interface," which creates three-dimensional video from video conferencing systems to encourage users to interact using natural gaze and body movement, "body movement generation technology," which automatically generates gestures based on the speaker's speech and language, and "personal characteristic/skill estimation technology," which accurately estimates the personal characteristics and communication skills of participants based on speech, language, and image information from their conversation.

—What are the current challenges you're facing?

One useful approach is applying machine learning technology, which is already widely studied and used, to the field of multimodal interactions. This can help us understand and model how humans' intentions are conveyed, and the mechanisms of how we

transmit information to each other. Machine learning technology generally requires a large amount of data anyway, but when dealing with human communication, there's a massive diversity of communication methods due to the large number of variables, such as different situations, number of people, cultures, relationships, and locations. This makes collecting the necessary data a difficult task that requires a huge amount of work.

For example, when gathering data from a conversation, the first step is to record the content of the speech by listening to it and manually writing up the timing and what was said. For example, "XX was said between XX seconds and XX seconds" (although speech recognition technology can be used for preprocessing). To some extent, we can automatically gather data on facial expressions, gaze, and posture from images of the speakers. However, for some modalities this method isn't accurate enough, so each image must be checked manually, by checking who's looking at who and labeling the eye movement, for example. A one-second video usually consists of about 30 still images, which makes the manual process time consuming and cumbersome. For these reasons, it takes a lot of work to construct corpus data (in this case, a data group for dialogue research) containing multimodal information for human conversations, and in many cases, we're unable to collect all of the data and we have to conduct our research using only a small amount of data from a particular situation.

Researchers nowadays say that once the data are collected, the research is 70% or 80% done. It may be a bit of an exaggeration, but collecting data is a very important and very expensive task. Being able to efficiently collect large amounts of data from human conversation is a major challenge, but if it is overcome, I think research in this field will take a huge leap forward.

Aiming for an all-encompassing model for communication

—What are the plans for future research?

As a new approach, we are carrying out research into modeling all aspects of human communication in order to gain a deep understanding of the communication mechanisms. Current communication modeling technology defines a human behavior as input X and a single state such as a human emotion, a personal characteristic, or a location as output Y, and only works to understand the relationship between the two. These individual studies are very enlightening and interesting, but from the perspective of understanding and modeling the entire mechanism of human communication, they only cover a tiny number of phenomena. I have started to work on forming an integrated understanding and model of the relationships between the various phenomena that occur in human communication.

As an example, a person may express (or "transmit") multimodal information such as speech, language, and body movement through their actual behavior, but before they get to that point, they will have a certain internal mindset and intention. People will also have individual characteristics, such as a personality and values. In addition, the dialogue will be influenced by personal relationships, roles, and the atmosphere during the conversation. **Figure 2** shows a simple example of this relationship. We have modeled the entirety of natural communication in four different layers: a "high-order layer" that holds people's relationships, roles, and the atmosphere of the conversation; an "actual-behavior layer" for multimodal information transmitted and received through actual behaviors such as speech, language, and body movement; a "mental layer" that holds internal information such as people's mindsets and intentions; and a "personal-characteristics layer" that holds information such as personality and values.

As mentioned above, there have been many studies carried out in the past on things like using one behavior to predict the next, and estimating individual personal characteristics and relationships from behavior, but essentially, communication can be thought of as a model in which the four layers are interrelated and act in combination with each other. In addition, Fig. 2 shows the communication status at one point in time, but this model will change over time.

The current goal is to create an ultimate, all-encompassing model of communication by looking at the relationships between these layers and the changes over time.

—What kinds of possibilities will this technology unlock?

If we can create this ultimate, all-encompassing model of communication, the system will not only be able to accurately understand the state of the conversation, but will also be able to predict and simulate the future state. We expect this to enable three main things.

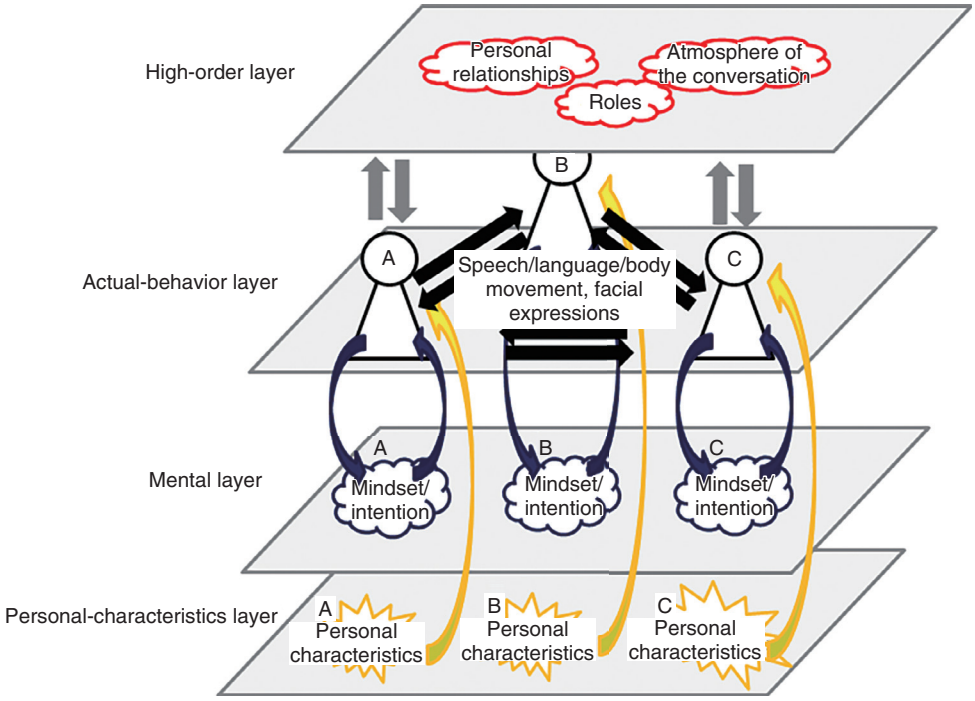

Fig. 2. Communication model devised by Dr. Ishii.

The first is the provision of real-time support for human communication. During a conversation, for example, if someone is a little upset, the system can follow up with them, or if one person is talking all the time it can ask what others think to try and switch the speaker and facilitate conversation, making the situation more harmonious and encouraging communication.

The second is the creation of the ultimate dialogue system. This dialogue system will be able to understand its conversation partner and the communication situation, and communicate appropriately just like a person.

The third is training people's communication skills. We are currently working on research into estimating how well we compliment people and automating training for this skill. Complimenting people is an important skill, and many people are concerned about how to do it best. With the all-encompassing communication model, it will be possible to evaluate how effective a person is at complimenting their conversation partner based on a wide range of perspectives such as their mindset and intention, their personal characteristics, the circumstances of the conversation, and the personal characteristics of their partner. Based on these evaluations, the system will give advice like, "A conversation partner with XX personality has XX intentions in XX situation, so complimenting XX more will help you build a better relationship." The aim is to use this ability to create applications that can improve people's communication skills.

—Do you have any messages for people aiming to become researchers?

What I'm aiming toward and what I want to achieve hasn't changed much since I was a student. I think my current research topic is my life's work, and I'm pretty invested in it. I think it's really important to spend your life in pursuit of what you want to achieve, and not to give up. There are opportunities for research and development in universities, companies, and various other places, but I don't think there's any environment where you will be 100% satisfied. At times, you may have to work on research topics that don't match your own aims. I think it's a real waste if you let yourself get dissatisfied in that kind of situation, and lose your motivation for research and your appetite for developing your personal skills. Whatever situation you find yourself in, you should think hard about what you can learn to help you reach your

final goal, how to achieve results efficiently, and how to make the most of your own abilities to succeed. And I think it's very important to build up your achievements one by one. In doing so, I believe that opportunities to pursue the research topics you're interested in will come around, and you will also have opportunities to start up the research yourself.

I also think it's important to involve other people. I am currently conducting four joint research projects with researchers from outside the company, and I have lots of other talented colleagues outside the company who I can work together with to help further my research. There are real limits to what you can achieve on your own. I think it's very important that you build up a group of colleagues who can work together to help each other achieve their goals.

I think multimodal interaction is a great research area for those who like people and are interested in communication. Multimodal interaction is an interdisciplinary field that involves a wide range of academic areas, including humanities and engineering. That means a single area of expertise alone is not enough for the creation of good technology, and a wide range of expertise is required. It's not easy to gain expertise across such a broad area, and this is still a new field that's under development. In order to break new ground in this field and succeed as a researcher, you need a strong desire to understand the mechanisms of human communication and to change the world through these new interaction technologies. You may not have the necessary expertise at the start, but I think if you have a strong drive to get stuck into the research, you'll be able to succeed in this field.

■ **Interviewee profile**
Ryo Ishii
Distinguished Researcher, NTT Human Informatics Laboratories.

He received a B.S. and M.S. in computer and information sciences from Tokyo University of Agriculture and Technology in 2006 and 2008, and a Ph.D. in informatics from Kyoto University in 2013. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 2008, he worked for NTT Cyber Space Laboratories (2008–2012), NTT Communication Science Laboratories (2012–2016), NTT Media Intelligence Laboratories (2016–2021, including as a visiting researcher at Carnegie Mellon University from 2019–2020). He has been with NTT Human Informatics Laboratories since July 2021 and with NTT Digital Twin Computing Research Center since January 2021.