Thought-processing Technology for Understanding, Reproducing, and Extending Human Thinking Ability

Kyosuke Nishida, Takeshi Kurashima, Noboru Miyazaki, Hiroyuki Toda, and Shuichi Nishioka

Abstract

The Human Insight Laboratory at NTT Human Informatics Laboratories aims to deepen our understanding of human thinking ability, which consists of human information understanding, human information processing, and response and behavior, and to create technologies that can reproduce these abilities on computers and expand human thinking ability. In this article, the technologies of visual machine reading comprehension, behavior modeling, speech recognition, and thought-enhancing stimulus design are introduced.

Keywords: human information understanding, human information processing, human insight

1. Introduction

The goal of Human Insight Laboratory of NTT Human Informatics Laboratories is to deepen our understanding of (1) human information understanding (from perception to cognition), (2) human information processing (for interpreting acquired information and taking the next action), and (3) reaction and action (for acting on the outside world). We create technologies to reproduce these processes on computers for expanding human thinking ability. We introduce four technologies that will lead to this goal, i.e., (i) visual-machine-reading-comprehension technology for understanding documents visually, (ii) behavior-modeling technology for considering the mechanism of human decision-making and behavior, (iii) speech-recognition technology for understanding a human's internal state, and (iv) thought-enhancing stimulus design for drawing out and expanding a human thinking ability.

2. Visual-machine-reading-comprehension technology for understanding documents visually

To further develop information retrieval and dialogue/question-answering services, the Human Insight Laboratory has been researching machine reading comprehension, by which artificial intelligence (AI) reads text written in natural language (i.e., the language used in our daily lives) and understands its meaning [1, 2].

Research on machine reading comprehension has made great progress, and according to evaluation data, AI has exceeded human reading comprehension; however, machine reading comprehension is limited in that it can only understand textual information. The PDF (portable document format) documents and presentation slides that we use on a daily basis contain not only linguistic information but also visual elements such as font size and color, figures, tables, graphs, and layout information. An integrated understanding of vision and language is therefore essential in regard to developing AI to support office work and daily life.

To achieve visual machine reading comprehension

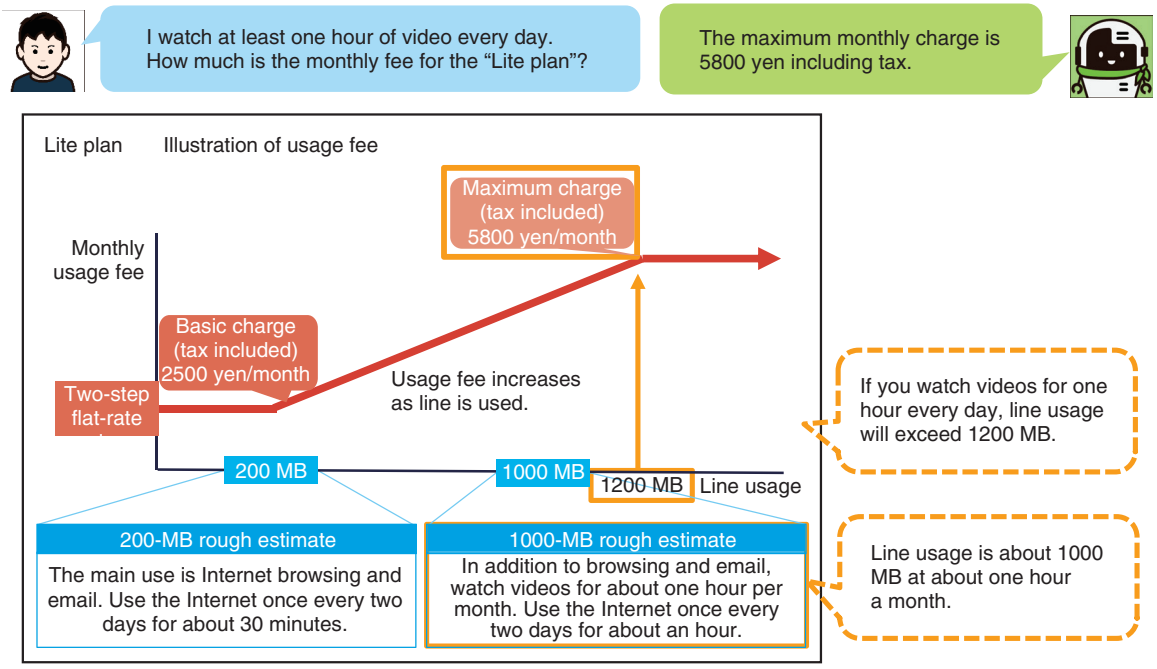


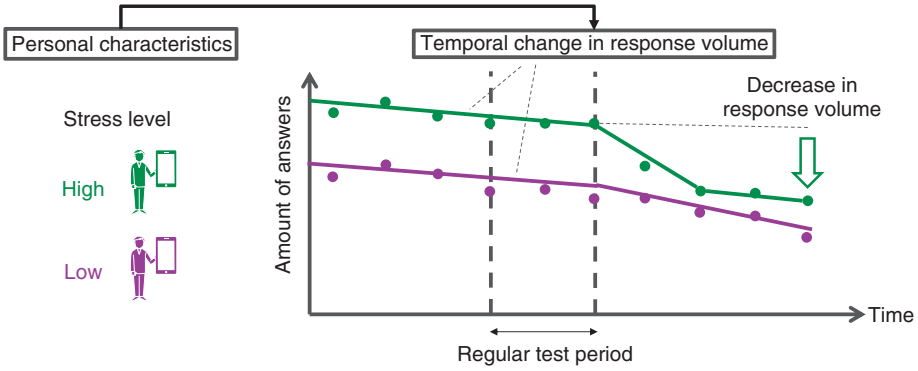
Fig. 1. Visual machine reading comprehension.

of document images (**Fig. 1**), we at the Human Insight Laboratory have constructed a dataset called VisualMRC [3] and are pushing ahead with research using it. This dataset consists of free-text question-answer data concerning document images of webpage screenshots, and regions in documents are annotated with nine classes, including title, paragraph, list, image, and caption. We conducted research using this dataset and proposed a visual-machine-reading-comprehension model that can take into account—as additional input—areas in the document (extracted using object-recognition technology) as well as position and appearance information of “tokens” (extracted using optical character recognition technology) [3]. Although this model is not yet able to match the accuracy of human question answering, we confirmed that understanding the visual information of documents improves the performance of question answering compared with that of text-only models. We will continue to work on various research projects for integrated understanding of vision and language.

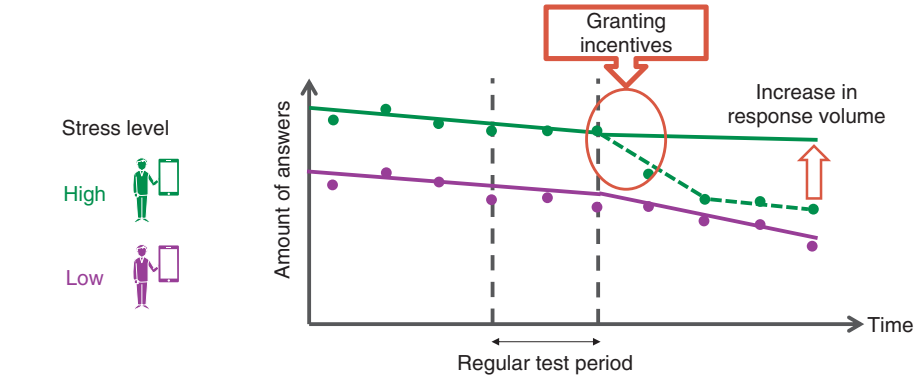
3. Behavior-modeling technology for considering the mechanism of human decision-making and behavior

We are studying the mechanisms of human decision-making and behavioral judgments (including irrational judgments) by using both humanistic and sociological knowledge (such as behavioral economics) and a variety of human data that have become available with the recent spread of the Internet of Things. We are also constructing a behavioral model that reproduces human decision-making and behavior by using the obtained knowledge as a framework and aim to apply this model to enable future predictions and simulations about people. When an undesirable future is predicted, we will be able to use the simulation results to search for ways that will lead to a brighter future.

As part of our efforts to explain the mechanism of human decision-making, we have been studying the effects of factors, such as personality, values, physical and psychological states, and social environment, on a person’s decision-making and behavior. Ecological momentary assessment (EMA), also known as the experience sampling method, is a survey method with which participants are sent simple questions about their situation, thoughts, emotions, and behavior via a portable device (such as a smartphone or tablet) and asked to spontaneously respond to the questions. Using the results of a 10-week EMA questionnaire administered to university students [4], we analyzed in detail the differences in the students’



(a) Analysis of temporal changes in the relationship between personal characteristics (stress level), social environment (regular examination), and response volume of EMA

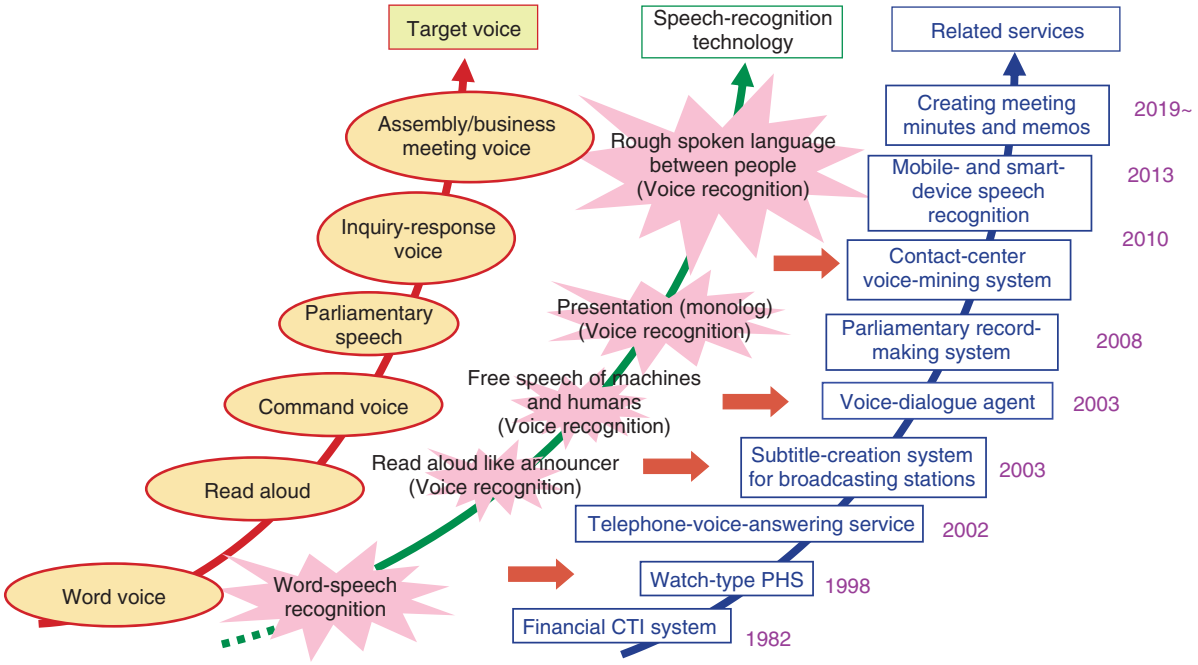


(b) Give an incentive when the amount of responses decreases and encourage an increase in the amount of responses

Fig. 2. Search for strategies on the basis of results of behavioral-data analysis.

tendency to respond to the questionnaire on individual characteristics from the perspective of temporal changes [5]. For example, as shown in **Fig. 2(a)**, participants who were judged to have high stress levels on a daily basis (according to a preliminary survey taken before the start of the EMA questionnaire) tended to actively self-disclose information to their device throughout the survey period. However, our analysis revealed that immediately after a stressful event (e.g., a regular exam), the students tended to be uncooperative toward answering, and their volume of responses dropped sharply. We also found that the participants who were judged to have high integrity in the preliminary survey were cooperative and gave a large amount of responses immediately after the start of the EMA questionnaire; however, the amount of their responses tended to decrease significantly over time. By deepening our understanding of the temporal changes in such human behavior (in this case, the act of voluntarily answering a questionnaire), it will be possible to predict what will happen regarding people's behavior and outcomes, and it will be easier to make decisions on how to change the future for the better. From the viewpoint of constantly understanding a person's condition, the EMA questionnaire should secure a uniform number of responses without bias in time or participant. For example, when the amount of responses is expected to drop, as shown in **Fig. 2(b)**, it is possible to give special incentives to participants with high stress levels to encourage them to respond after the test.

The above example is just an analysis that captures one aspect of complex human behavior. In the future, we plan to make our human decision-making and behavioral models [6] more sophisticated by taking into account behavioral economic human characteristics



CTI: computer telephony integration
PHS: personal handy-phone system

Fig. 3. Efforts concerning voice-recognition technology at NTT.

such as attitude toward uncertainty (risk-aversion tendency) and toward waiting (patience). By "sophisticated," we mean that such a model will be able to make more "human-like" decisions. We believe that because a model is "human-like," it will be possible to predict and depict a future of society, even an undesirable future.

4. Speech-recognition technology for understanding a human's internal state

With the advent of voice assistants, which allow users to operate devices by speaking into a smartphone or AI speaker, speech-recognition technology has spread rapidly around the world. The practical application of speech-recognition technology for operating computers by means of spoken dialogues using short-word exchanges began in the 1980s with the introduction of interactive voice-response systems. The introductions of car (satellite) navigation systems in the 1990s and deep-learning technology in recent years have greatly improved speech-recognition accuracy and spurred the rapid spread of voice assistants.

Natural communication between people involves longer sentences and more complex linguistic expressions, which are more difficult for speech recognition to handle, than the short sentences uttered to a voice assistant. Moreover, when the conversational partner is familiar, the tone of voice becomes more casual and utterances fragmented, which makes it more difficult to predict both sounds and linguistic expressions. Regardless of those difficulties, recognizing such conversational speech can be used for applications with significant business value, such as analyzing the content of conversations in call logs stored in large quantity at call centers and supporting conversations in real time; accordingly, various speech-recognition technologies based on deep learning are being actively studied. As shown in **Fig. 3**, speech-recognition technology has evolved through repeated improvements in the accuracy at which speech can be converted to text, expansion of its applications, and increasing complexity of the speech to be handled.

Information communicated through speech includes not only verbal information (text information) but also non-verbal information such as gender, age, emotions, intentions, and attitudes. As well as working on recognizing text information from speech with high accuracy, we are investigating technologies for

recognizing and using non-verbal information and developing technologies that can extract speaker attributes (adult male, adult female, and child), emotions (joy, anger, sadness, and calmness), and questions and/or non-questions. We are also researching, developing, and practically applying technologies for estimating the degree of a customer's anger and/or satisfaction as well as the impression that the telephone operator's response gives to that customer during conversational speech consisting of two speakers (such as when a customer calls a call center).

The recognition technology for verbal and non-verbal information we are currently investigating is the first step toward reading the internal state of humans in a manner that allows us to offer more-advanced services. In everyday conversation, we use various clues, such as voice tone, line of sight, facial expression, pause length during interaction, and changes in wording, to estimate information inside a person's mind, such as their emotions and interests or friendliness and indifference, and we use that information to facilitate smooth communication. The recognition of such internal information using a machine is expected to lead to new services that target the subtleties of the human mind that cannot be expressed in words. Such services might conceivably include a voice-dialogue agent that is attuned to the emotions of the user, education that responds to the understanding and interests of each student, and detection of a patient's physical condition and mental stress from medical-interview dialogue data. It is also expected to contribute to creating a new type of communication—as targeted by the grand challenge "Mind-to-Mind Communications" set for Digital Twin Computing (DTC), which is one of the elements of IOWN (the Innovative Optical and Wireless Network) promoted by the NTT Group, that can directly understand the way people perceive and feel in a manner that transcends differences in individual characteristics such as experience and sensitivity.

5. Thought-enhancing stimulus design for drawing out and expanding a human thinking ability

Human thoughts and behaviors are not only influenced by information that is interpreted consciously, such as language and numerical values, but also by sensations and perceptions that are perceived unconsciously, such as colors, scents, and the atmosphere created by the manner of speaking. For example, smelling soap makes us want to wash [7] and changing the color of a product package makes us feel that the product is different even though it contains the same thing [8].

We are accumulating knowledge on the relationship between perceptual stimuli and human thought and behavior, and on the basis of that relationship, we are researching generation and control of perceptual stimuli with the aim of drawing out and expanding human thinking ability. In one of our previous studies, we investigated the effect of the speaker's speaking style (voice pitch, speaking speed, and inflection) on a human's psychological state and behavior [9]. In that study, we conducted a large-scale subjective-evaluation experiment on the relationship between speech that promotes products and people's consumption behavior and analyzed the relationship between speech and purchase intention by using the consumer-behavior model developed by Mehrabian and Russell [10] (**Fig. 4**) that expresses emotion as a mediator.

5.1 Experimental procedure

Via crowdsourcing, we asked 202 native speakers (male and female) of Japanese to listen to an advertisement for an electrical appliance spoken in different styles of speech. After listening, we asked them to respond to evaluation items about the emotion and purchase motivation they felt when listening to the spoken ad. We then analyzed the relationship among speech characteristics, emotions, and purchase intention on the basis of the obtained parameters (voice pitch, speaking speed, and intonation) related to the ad's speech (voice characteristics) and the responses to the evaluation items. The promotional text of the ad was based on speech uttered by a professional female speaker in a read-aloud tone, where the speech parameters were set as shown in **Table 1**. The evaluation items are listed in **Table 2**.

5.2 Results

The results of analyzing the relationship among speech characteristics, emotions, and purchase intention—by using a three-layer model based on Mehrabian and Russell's consumer-behavior model—are shown in **Fig. 5**. In line with previous studies [11], the results indicate that among the emotions, "pleasure" and "arousal" positively affect purchase intention. They also indicate that the effect of pleasure is the strongest, higher voice, faster speech speed, and higher intonation lead to higher emotion rating, and among those parameters, speech speed had the greatest

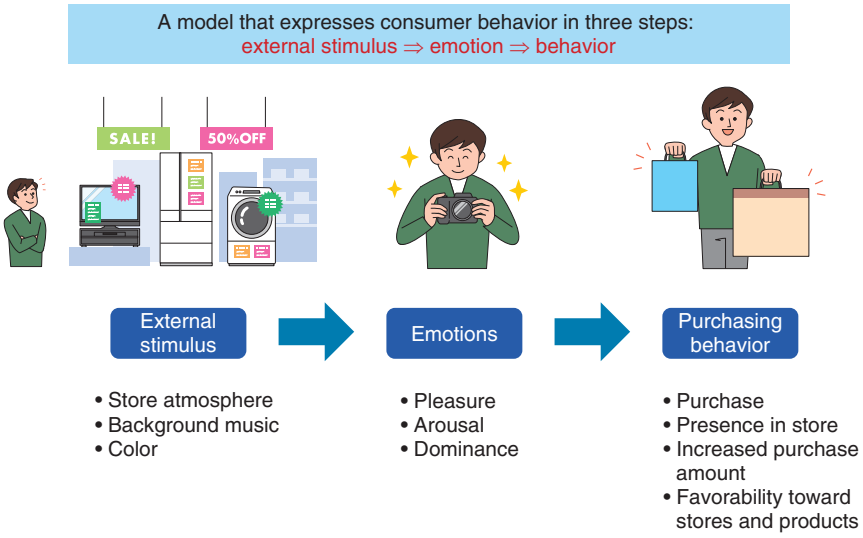


Fig. 4. Consumer-behavior model.

Table 1. Voice parameters.

Parameters of voice		Setting value
Voice pitch	Average F0 (Hz)	0.9439x, 1x, 1.059x
Speaking speed	Speaking speed (mora/s)	0.890x, 1x, 1.1225x
Magnitude of intonation	F0 variance (Hz)	0.6667x, 1x, 1.5x

Table 2. Evaluation items.

Evaluation item		Setting value
Emotion	Pleasure	Displeasure ⇔ Pleasure (7 stages)
	Arousal	Sleep ⇔ Arousal (7 stages)
	Dominance	Obedient ⇔ Dominant (7 levels)
Purchase intention		“I want to buy it very much.” ⇔ “I don’t want to buy it at all.” (7 levels)

effect on the evaluation items.

To summarize these results, we found that (i) the consumer-behavior model is valid for a speech stimulus based on read-out advertisements and (ii) the speed of speech has the strongest effect on pleasure, which has the strongest effect on purchase intention.

The above-described efforts represent the analysis of only a small part of the complex and diverse thinking processes of humans. We plan to study a wider range of factors, such as the relationship among the attributes of listeners with an auditory stimulus, and investigate non-auditory stimuli such as sight and smell. We also plan to study the effects of stimuli on human thinking phenomena—ranging from intuitive judgments to deeper stages such as effects on logical thinking and human values. Through these studies, we want to clarify what is necessary to draw out human thinking ability, enhance human potential, and contribute to the creation of a better society.

6. Concluding remarks

By refining the four technologies described in this article, we plan to understand, reproduce, and extend human thinking ability, which consists of human information understanding, human information

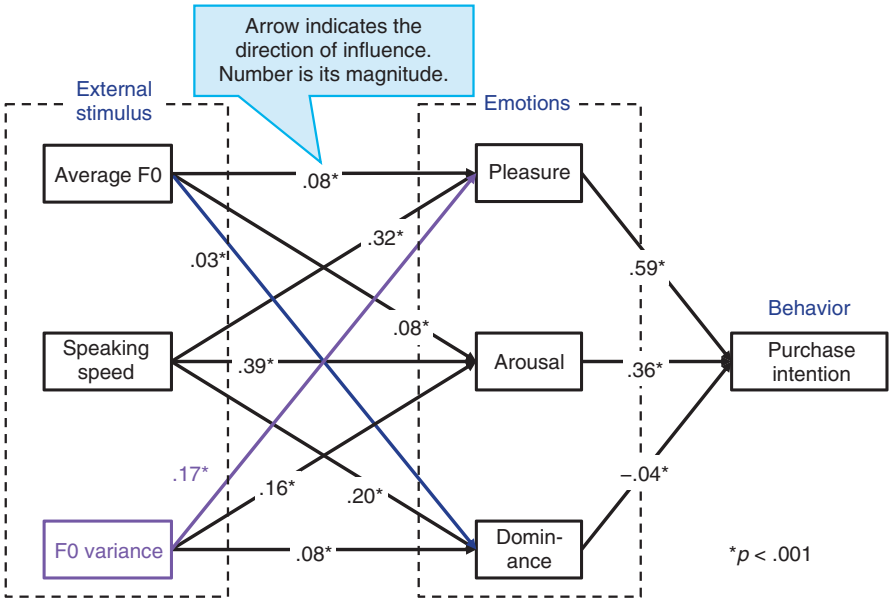


Fig. 5. Results of analysis of relationships among voice characteristics, emotions, and purchase intention.

processing, and response and behavior.

References

[1] K. Nishida, I. Saito, K. Nishida, K. Shinoda, A. Otsuka, H. Asano, and J. Tomita, "Multi-style Generative Reading Comprehension," Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), pp. 2273–2284, Florence, Italy, July 2019.

[2] K. Nishida, K. Nishida, M. Nagata, A. Otsuka, I. Saito, H. Asano, and J. Tomita, "Answering while Summarizing: Multi-task Learning for Multi-hop QA with Evidence Extraction," Proc. of ACL 2019, pp. 2335–2345, Florence, Italy, July 2019.

[3] R. Tanaka, K. Nishida, and S. Yoshida, "VisualMRC: Machine Reading Comprehension on Document Images," Proc. of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021), pp. 13878–13888, Feb. 2021.

[4] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones," Proc. of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2014), pp. 3–14, Seattle, USA, Sept. 2014.

[5] T. Tominaga, S. Yamamoto, T. Kurashima, and H. Toda, "Effects of Personal Characteristics on Temporal Response Patterns in Ecological Momentary Assessments," Proc. of the 18th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2021), Bari, Italy, Aug./Sept. 2021.

[6] T. Kurashima, T. Althoff, and J. Leskovec, "Modeling Interdependent and Periodic Real-world Action Sequences," Proc. of the 2018 World Wide Web Conference, Lyon, France, pp. 803–812, Apr. 2018.

[7] K. Liljenquist, C. B. Zhong, and A. D. Galinsky, "The Smell of Virtue: Clean Scents Promote Reciprocity and Charity," Psychological Science, Vol. 21, No. 3, pp. 381–383, Mar. 2010.

[8] K. Maki, "Color Studies for Color Design," Ohmsha, 2006 (in Japanese).

[9] M. Nagano, Y. Ijima, and S. Hiroya, "Impact of Emotional State on Estimation of Purchase Intention from Advertising Speech," Proc. of INTERSPEECH 2021, Brno, Czech Republic, Aug./Sept. 2021.

[10] A. Mehrabian and J. A. Russell, "Approach to Environmental Psychology," The MIT Press, 1974.

[11] R. Donovan and J. Rossiter, "Store Atmosphere: An Environmental Psychology Approach," Journal of Retailing, Vol. 58, No. 1, pp. 34–57, 1982.

Kyosuke Nishida
Distinguished Researcher, Human Insight Laboratory, NTT Human Informatics Laboratories.

He received a B.E., M.I.S., and Ph.D. in information science and technology from Hokkaido University in 2004, 2006, and 2008 and joined NTT in 2009. He received the Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) in 2015 and the Kambayashi Young Researcher Award from the Database Society of Japan (DBSJ) in 2017. His current interests include natural language processing and artificial intelligence. He is a member of the Association for Computing Machinery (ACM), IPSJ ,the Association for Natural Language Processing (NLP), the Institute of Electronics, Information and Communication Engineers (IEICE), and DBSJ.

Takeshi Kurashima
Distinguished Researcher, Human Insight Laboratory, NTT Human Informatics Laboratories.

He received a B.S. from Doshisha University, Kyoto, in 2004 and M.S. and Ph.D. in informatics from Kyoto University in 2006 and 2014. He joined NTT in 2006. He was a visiting scholar at Stanford University, USA from 2016 to 2017. His current research interests include data mining, machine learning, and recommender systems. He is a member of IPSJ, IEICE, DBSJ, and ACM.

Noboru Miyazaki
Senior Research Engineer, Supervisor, Human Insight Laboratory, NTT Human Informatics Laboratories.

He received a B.E. and M.E. from Tokyo Institute of Technology in 1995 and 1997. Since joining NTT in 1997, he has been engaged in research on spoken dialogue systems, speech synthesis, and speech recognition technologies. He received the IEICE Inose Award in 2001. He is a member of the Acoustical Society of Japan, IEICE, and the Japanese Society for Artificial Intelligence (JSAI).

Hiroyuki Toda
Senior Research Engineer, Supervisor, Human Insight Laboratory, NTT Human Informatics Laboratories.

He received a B.E and M.E. in materials science from Nagoya University in 1997 and 1999, and Ph.D. in computer science from University of Tsukuba in 2007. He joined NTT in 1999. His current research interests include information retrieval and data mining. He is a member of IPSJ, IEICE, JSAI, DBSJ, and ACM.

Shuichi Nishioka
Executive Reserch Engineer, Supervisor, Human Insight Laboratory, NTT Human Informatics Laboratories.

He received a B.E. and Ph.D. in engineering from Yokohama National University, Kanagawa, in 1995 and 2005. He joined NTT in 1995. His current interests include data engineering and artificial intelligence. He is a member of IPSJ and DBSJ.