Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 15, No. 1, March 2015, pp. 1-11
http://dx.doi.org/10.5391/IJFIS.2015.15.1.1

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

Deep Level Situation Understanding for Casual Communication in Humans-Robots Interaction

Yongkang Tang[1,2], Fangyan Dong[3], Yoichi Yamazaki[4], Takanori Shibata[5], and Kaoru Hirota[1]

[1]Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama 226-8502, Japan
[2]Applied Informatics, Faculty of Science and Engineering, Hosei University, Koganei 184-8584, Japan
[3]Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama 226-8502, Japan
[4]Department of Home Electronics, Faculty of Creative Engineering, Kanagawa Institute of Technology, Atsugi 243-0292, Japan
[5]Human Technology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8568, Japan

ijfis

Received: Jan. 9, 2015
Revised : Feb. 24, 2015
Accepted: Mar. 15, 2015

Correspondence to: Kaoru Hirota
(hirota@hrt.dis.titech.ac.jp)

Abstract

A concept of Deep Level Situation Understanding is proposed to realize human-like natural communication (called casual communication) among multi-agent (e.g., humans and robots/machines), where the deep level situation understanding consists of surface level understanding (such as gesture/posture understanding, facial expression understanding, speech/voice understanding), emotion understanding, intention understanding, and atmosphere understanding by applying customized knowledge of each agent and by taking considerations of thoughtfulness. The proposal aims to reduce burden of humans in humans-robots interaction, so as to realize harmonious communication by excluding unnecessary troubles or misunderstandings among agents, and finally helps to create a peaceful, happy, and prosperous humans-robots society. A simulated experiment is carried out to validate the deep level situation understanding system on a scenario where meeting-room reservation is done between a human employee and a secretary-robot. The proposed deep level situation understanding system aims to be applied in service robot systems for smoothing the communication and avoiding misunderstanding among agents.

Keywords: Human robot interaction, Multi-agent, Casual communication

1. Introduction

Robots are increasingly capable of co-existing with humans in environment, such as in manufacturing, offices, restaurants, hospitals, elder care facilities, and homes. The ability of comprehending human activities, e.g., gesture/posture, speech, and emotion, is required for robots in casual communication, i.e., human-like natural communication. Verbal and non-verbal communications are the two basic ways for transmitting messages among various agents such as humans and robots/machines in casual communication. Several spoken dialog systems are proposed for verbal communication [1, 2]. As for nonverbal approaches, gesture

recognition has become an attractive research theme in the field of robot control [3] and sign language recognition [4]. Most works on gesture recognition for Human-Robot Interaction (HRI) have been done based on visual information, such as sign tracking and recognition system (STARS) [5]. To improve the robustness of gesture recognition system, a Choquet integral based multimodal gesture recognition system [6] is proposed. Emotions and intentions are also important in HRI. An automatic real-time capable continual facial expression recognition system [7] is proposed based on Active Appearance Models (AAMs) and Support Vector Machines (SVMs), in which face images are categorized into seven emotion states (neutral, happy, sad, disgust, surprise, fear, and anger). An individual mean face is estimated over time to reduce the influence of individual features. A maximum entropy based intention understanding method [8] is proposed for understanding the intention of speech in a dialog system. In communication among multiple agents, e.g., a conference with twenty participants, it may not be easy to identify the attitude, mood, and emotion of each individual. A concept of Fuzzy Atmosfield (FA) [9] is proposed to represent the atmosphere being created in the process of interactive communication.

Human may hide their real emotions and intentions in casual communications. But other humans may be able to understand them to some extent by understanding the spoken contents, voice tones, and facial expression changes. Robots are also expected to be competent to these kinds of nature communications. Although speech recognition, gesture/posture recognition, emotion recognition, intention estimation, and atmosphere estimation can help robot to comprehend parts of human activities, these approaches are still insufficient for understanding the inner emotions and intentions of interlocutors in casual HRI. The agent dependent customized knowledge, e.g., normal state and habits information, should be considered. The audible information (e.g., speech and voice) and visible information (e.g., gesture, posture, and facial expression) are called surface level communication in this paper, while deep level situation understanding is characterized as unifying the surface level understanding, emotion understanding, intention understanding, and atmosphere understanding, thoughtfulness inference, and both universal and agent dependent customized knowledge for casual HRI. The deep level situation understanding framework consists of a gesture/posture recognition module, speech/voice recognition module, emotion recognition module, intention estimation module, atmosphere understanding module, and knowledge base (including universal knowledge and customized agent-dependent knowledge).

The deep level situation understanding in casual communication among various agents, e.g., humans and robots/machines, aims at three issues. Firstly, humans must pay special attention to robots in the ordinary human-machine communication systems, but such burden may be reduced if robots have deep level situation understanding abilities. Secondly, in the real world, unnecessary troubles or misunderstandings in human to human communications may sometimes happen but the deep level situation understanding can make it possible to avoid such lower level troubles. The customized agent-dependent knowledge will help to comprehend and avoid misunderstanding. Thirdly, with the consideration of surface level information, emotions, intentions, atmospheres, universal knowledge, and customized agent-dependent knowledge, it will also help to understand the background, habits, and intention of the agent for smoothing natural HRI, so as to create a peaceful, happy, and prosperous society which consists of humans and various specification robots/machines. The main contribution is the concept of deep level situation understanding which aims to realize human like deep level communication in HRI.

A simulated experiment is established to implement the proposed deep level situation understanding system where meeting-room reservation in a company is done between a human employee and a secretary-robot. The inference system is based on the hypothesis that the secretary-robot already accumulates thoughtfulness and customized knowledge during daily communication with employees. Twelve subjects are asked by questionnaires to evaluate the response of the proposed inference system comparing to the responses from familiar people.

The concept of deep level situation understanding is proposed in Section 2. In Section 3, an inference system of deep level situation understanding is constructed. A simulation experiment is carried out to evaluate the availability of the proposed inference system in Section 4.

2. Deep Level Situation Understanding

2.1 Concept of Deep Level Situation Understanding

Although speech understanding, gesture/posture understanding, emotion understanding, intention understanding, and atmosphere understanding can help robot to comprehend parts of human activities. These approaches are still difficult to understand human activity deeply in casual HRI. People usually hide their real emotions, intentions, and opinions and show them in another indirect/different way. These kinds of information are

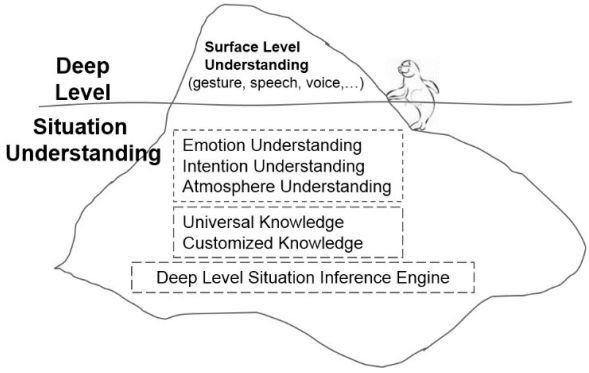

Figure 1. The relationship between the surface level understanding and the deep level situation understanding.

just a reflection of the real emotions, intentions, and feelings.

The audible information (e.g., speech and voice) and visible information (e.g., gesture, posture, and facial expression) are just the surface information of humans. Thus the understanding of such surface information is called surface level understanding in this paper. If the understanding level is illustrated as an iceberg, the audible and visible information is just like a tip of the whole iceberg above the sea level, while there still remains more information hidden under the sea level such as emotion, intention, and atmosphere. In contrast with surface level understanding, the deep level situation understanding is characterized as unifying the surface level understanding, emotion understanding, intention understanding, thoughtfulness inference, and both universal knowledge and agent dependent customized knowledge for casual HRI. The relationship between the surface level understanding and the deep level situation understanding is illustrated in Figure 1.

Moreover, customized knowledge and thoughtfulness should also be considered for casual humans robots communication. Customized knowledge and thoughtfulness are detailed in Section 2.1.1 and Section 2.1.2 respectively.

2.1.1 Customized knowledge

Why the communications between friends are usually smoother than the communications between strangers? It's because friends usually know each other very well. Friends have special knowledge, e.g., tempers, habits, and means of expression, of each other. These special knwoledges may help to avoid misunderstanding in human-human communications.

These special customized knowledge should also be considered in the humans-robots communication for realizing the smooth communications. There are two kinds of customized knowledge data. (1) The data that characterising the normal state (including the normal tones, normal facial expression) of a people. Because people may show their pleasure and anger in different ways. Some people may keep smiling face all days. When angry, they may just keep silent. For these people, smiling is their nomal state. (2) The data featuring poeple's habits (e.g., his/her like, frequency of doing something). The habit data is obtained from the history communications. Usually these kinds of data is known to freinds.

2.1.2 Thoughtfulness communications

Human usually consider emotion and intention of their conversation partners. There are many instances of deep level situation understanding in daily life. Suppose you visit a convenience store and want to buy a fountain pen. In this case, you may ask the shop assistant that "Do you have a fountain pen?" The shop assistant will know that you want to buy this kind of pen. Even if it is a yes-no question, neither "yes" nor "no" is expected to end the conversation. If there are fountain pens in the shop, the shop assistant will guide the customer to the specific location of the fountain pens. If they do not have this kind of pen, in order to provide satisfactory service to the customer, they may tell the customer where the fountain pen can be purchased. Another example, imagine a lady usually goes to a cafe to have her favorite coffee and dessert. The waiter/waitress in the cafe knows the preference of their regular customer. When this lady just orders "the usual one" it is no doubt that the waiter/waitress will understand the meaning and bring the desired drink and dessert to her.

Because thoughtful communication can make the conversation partners feel comfortable, the robots should also have the ability of thoughtfulness inference and act like humans.

2.2 Inference Framework for Deep Level Situation Understanding

Since people usually hide their real emotions, intentions, and opinions and show them in another indirect/different way. Not only the surface level information, i.e. visible and audible information, is important for human-robot communication, but also emotion information, intention information, and atmosphere information should be considered for human-like reactions.

The illustration of multi-modal framework for deep level situation understanding is shown in Figure 2. The audible information and visible/tactile information are obtained by microphones

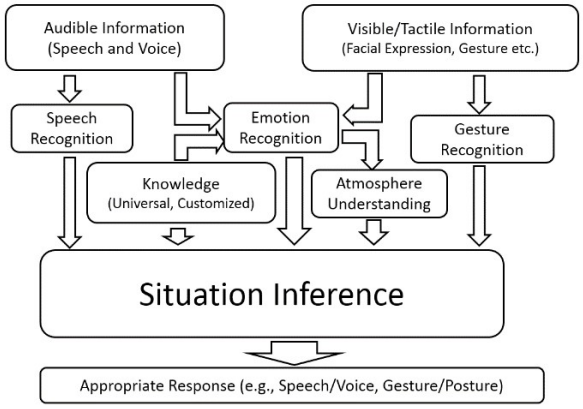

Figure 2. Multimodal framework for deep level situation understanding.

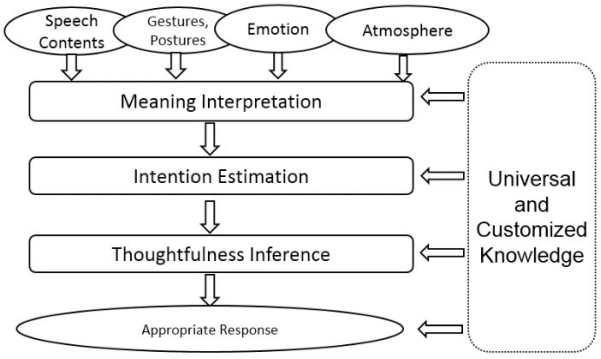

Figure 3. The flowchart of deep level situation inference.

and cameras/tactile sensors. The speech content is recognized by speech recognition method [10]. People may express their emotion in different ways. To estimate the real emotion of a people, the friend-level knowledge, i.e., customized knowledge of the people, is necessary. The face features and voice features of normal state is used to train the classifier. Then the real emotion state of the people is estimated by the trained classifier. Atmosphere is estimated based on the emotion state of the agents. Gestures/postures is recognized by gesture recognition algorithms [6] from sensors like cameras and accelerometers. The speech contents, universal and customized knowledge, emotions, atmospheres, and gestures/postures are important input for the situation inference system.

The inference flowchart of deep level situation understanding is shown in Figure 3. The meaning of the interlocutor is analyzed from the verbal information (speech contents) and the non-verbal information (gestures/postures). Then the intention is estimated based on the analyzed meaning and knowledge

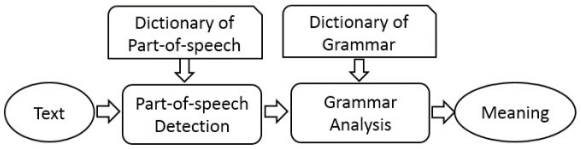

Figure 4. The flowchart of meaning analysis.

from historical dialogs. Thoughtfulness is inferred based on the intention and the thoughtfulness knowledge. Finally with the comprehensive consideration of the current emotion state, atmosphere, universal and customized agent-dependent knowledge, and thoughtfulness, suitable response (speech, voice, and gesture/posture) is reasoned and then outputted as the final result of the system. The details of situation inference is mentioned in Section 3.

3. Situation Inference System

The proposal is a part of the project called "Multi-Agent Fuzzy Atmosfield", which contains several on-going research themes, such as research of deep level situation understanding, deep level emotion understanding based on customized knowledge [11], and atmosphere understanding for HRI [12]. The deep level emotion understanding method has been proposed for agent to agent communication, where customized learning knowledge of an observed agent is used with the observed input information from Kinect. The aim of the deep level emotion understanding is to realize agent dependent emotion understanding by utilizing special customized knowledge of the agent, rather than ordinary surface level emotion understanding by using visual/acoustic/distance information without any customized knowledge. The proposal mainly focuses on inferring based on the text of utterance.

3.1 Meaning Interpretation

Verbal and non-verbal communications are two natural ways in humans-robots communications. The verbal information (i.e., speech) is able to be transferred into text sentence by means of speech recognition library (e.g., Julius [10] for Japanese speech recognition). The non-verbal information (i.e., gesture and posture) is recognized by gesture recognition algorithms [5].

The flowchart of meaning analysis is shown in Figure 4. Firstly, the text of utterance is divided into words list with

Table 1. Meeting room information stored in knowledge base

Knowledge_id	Property	Value	Value_ toplimit
room1701	type	meetingroom	
room1701	room-no	1701	
room1701	floor	17	
room1701	capacity	10	
room1701	condition	quiet	
room1701	available-time	14:00	18:00

part-of-speech tags. The boundary of Japanese utterance is determined by a conditional random field method [13].

Secondly, a dictionary of grammar is employed to transfer the words list into meaning string. For example, the utterance "Is there a meeting-room available from 15:00 PM?" is converted to the meaning string, "Query (subject=meetingroom, available-time≤15:00)".

After got the meaning of utterances, it is easy to transfer the meaning string into a Structured Query Language (SQL) statement and execute on the knowledge database. Table 1 shows an example of a meeting room stored in knowledge database. The corresponding SQL sentence of previous example will be "select distinct t0.knowledge_id from knowledge t0 where 1 = 1 and t0.knowledge_id in (select t1.knowledge_id from knowledge t1 where t1.knowledge_id = t0.knowledge_id and t1.property = 'type' and t1.value ='meetingroom') and t0.knowledge_id in (select t2.knowledge_id from knowledge t2 where t2.knowledge_id = t0.knowledge_id and t2.property = 'available-time' and strftime('%s', t2.value) <= strftime('%s', '15:00')) ".

3.2 Intention Understanding

The intention of utterances is able to be understood from the customized knowledge data which is extracted from the history data of communications. A general communication process between agent A and agent B is illustrated in Figure 5. For the question from agent A, agent B may reply with many kinds of responses (e.g., Response 1, Response 2 ... Response N). Agent A may intend different intentions with some frequency for each response of agent B. For example, agent A may intend to intention 1, intention 2, and intention 3 with a frequency of P1, P2, and P3. Suppose P2 and P1 are the biggest and second biggest among P1, P2, and P3. If the difference between P2 and P1 is significant, there is no doubt that agent A will purport intention 2 definitely. Agent B will respond to intention

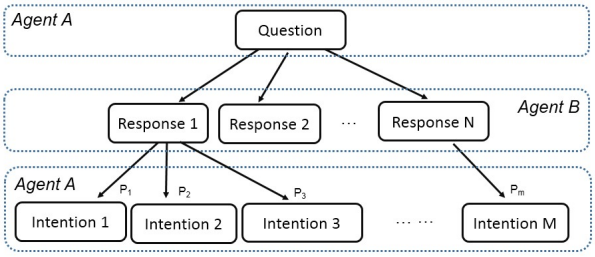

Figure 5. Illustration of conversation between agent A and agent B.

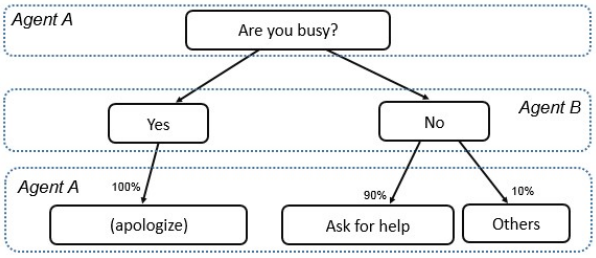

Figure 6. An example of communication between agent A and agent B.

2 directly. If the difference between P2 and P1 is tiny, that means agent A may purport either intention 1 or intention 2. Agent B may respond to the question from agent A by "Do you mean intention 2" because P2 is a little bigger than P1. Then agent B could respond with intention 2 directly when agent A asks this question.

Utterances usually contain some important information, e.g., people usually order their favorite food more frequently than the others. These kinds of habit information of a person are some kinds of deep level information which is only known by their friends. These information may be extracted from the utterances and accumulated as the customized knowledge.

The knowledge of intention is supposed to be collected by calculating the frequency of request and response from the long-term communication data between interlocutors. The simulation is carried out by using pre-established knowledge in this experiment for the sake of simplicity. An example is shown in Figure 6. Agent A asks agent B "Are you busy?" As illustrated in the figure, when agent B is busy and replies "Yes", agent A may respond "Sorry to bother you". When agent B is available and replies "No", the frequency of agent A asks for help is 90% and the frequency of agent A ask for others is 10%. In this situation, if agent B is available, then agent B almost definite that agent A intend to ask for help. After understand the intention

Table 2. Example of thoughtfulness knowledge

Intention	Condition	Response
Reserve room	No room available at specified time	Try to revere at other time
Reserve room	Specified room is not available	Try to reserve other room
Reserve room for remote conference	Only reserved meeting room	Reserve the video conference system
Reserve room for remote conference	Only reserved conference system	Reserve the meeting room

of agent A, the conversation between two agents will moves forward smoothly.

3.3 Thoughtfulness Inference

Thoughtfulness, i.e., showing kindly consideration for others, is a high level intelligent action of human which is usually performed between familiars.

The thoughtful response can be reasoned based on the intention and customized knowledge. Assuming in some company, the TV conference system should be reserved with the meeting room together when hold a video conference with branch companies. If an employee intends to have a TV conference with branch companies. But he just reserves the meeting room, it will be very helpful to remind him to reserve the TV conference system with the meeting room together.

Thoughtfulness knowledge is a known common knowledge of human. For robots, thoughtfulness response may be inferred based on the estimated intention and customized thoughtfulness knowledgebase. An example of thoughtfulness knowledge manually generated by trial and error is shown in Table 2. Accumulating the thoughtfulness knowledge from the communication data, however, should be studied in the succeeding research of this project.

Based on the result inferred from previous processing, expert rules are used to reason the reply utterance. In the conversation between an employee and a secretary-robot, if the employee just asks "Are you busy?" The secretary robot guesses that he is intended to reserve meeting-room. Reading the situation is more important for the secretary robot based on the emotion state of the human employee rather than expressing the emotion of the secretary robot to the human employee, because robot is supposed to have ability to adjust human characteristics in the robot-human co-existing society. If the emotion state of the employee is as usual, the response from the secretary-robot may be "Are you going to reserve a meeting-room?". if the employee looks sad and abnormal, the secretary-robot may respond to his latent request (intention) directly by "Please!". An example of expert rules are shown in Table 3.

Table 3. An example of expert rules

if query_ availability _of_robot then

if provability_of_intending_reserve_room ≥ 0.9

and emotion_state $=$ *normal then*

return "Are you going to reserve a meeting-room ?"

else if provability_of_intending_reserve_room ≥ 0.9

and emotion_state $=$ *abnormal then*

return "Yes, Please!"

else

return "What can I do for you?"

end if

end if

4. Experiment for Deep Level Situation Understanding

4.1 Experiment Setting

A simulation experiment is carried out to evaluate the naturalness of response inferred from proposed deep level situation understanding mechanism. A company scene is taken into consideration. There is a secretary robot who is supposed to do clerical works such as the reservation of hotel, meeting room, and TV conference system. One employee usually goes to secretary room to reserve meeting room. Assume the secretary robot has the following knowledge about this employee:

1) This employee usually comes to the secretary room for some help. When he asks the secretary "Are you busy?", the probability of asking to reserve meeting room is 0.9; the probability of asking for other help is 0.1;

2) Meeting room and TV conference system should be re-

Table 4. Script of reserving a meeting room

Employee*: Are you busy?*
[O1] Secretary-robot: *No, would you like to reserve a room?*
Employee: *Is the meeting room for 10 people vacant at 3 o'clock this Thursday.*
[O2] Secretary-robot*: They are available from 15:30.*
Employee: *Great! A quiet room is preferable.*
[O3] Secretary-robot: *How about the regular conference room on the 17th floor?*
Employee: *Sounds good! It's for a remote conference with the branch office, please reserve it until 17 o'clock.*
[O4] Secretary-robot*: In addition, I will reserve the video conference system, too*
Employee: Thanks!
Secretary-robot: You are welcome.

Table 5. Script of changing the schedule of meeting

Employee*: [Looks Sad]Are you busy now?*
[O5] Secretary-robot*: Go ahead!*
Employee*: Excuse me, can you change the meeting with the branch office to Thursday next week?*
[O6] Secretary-robot*: Sure! I will check it now.*
Employee*: Yes, please.*
[O7] Secretary-robot*: For Thursday next week, all the conference rooms on 17th floor have been scheduled already. How about the 11th floor conference room?*
Employee*: Great! I feel relieved.*
[O8] Secretary-robot*: I will also update the reservation of the remote conference system.*
Employee*: Thank you very much.*
Secretary-robot*: You're welcome.*

served together for the purpose of holding a remote meeting with the branch companies.

In the scenario, the employee comes to the secretary room to reserve meeting room. When he asks "Are you available now", the secretary is aware of that he is intent to reserve a meeting room. When the meeting room he wanted is not available at that time, the secretary checks other available time and recommends his favorite to him. When he tells the secretary, he wants to reserve for a remote conference, the secretary also reserves the TV conference system because it is necessary for holding a remote conference. Again, when he looks sad and wants to change the schedule of the reserved room, the secretary responds immediately to calm down the employee. When the room he wanted is not available, the secretary also checks the other rooms instead of responding as "No available room" directly. Finally, the secretary also change the schedule of the remote system for the employee.

The experiment system is developed in C++ language. All of the knowledge are saved to a SQLite database. This research which is carried out in text of Japanese mainly focuses on inferring based on the text information. Mecab is used to detect part-of-speech for Japanese [13]. The utterances of the employee are used as the input of the proposed inference system. The response of the secretary-robot is inferred as the output of the proposed inference system. The translated scripts of reserving a meeting-room and changing the schedule of reserved meeting-room are shown in Tables 4 and 5.

Rule generation is manually done by trial and error thinking about keeping the accuracy of intention estimation. After that, the naturalness of predetermined scenarios is investigated by doing a questionnaire to evaluate the naturalness of the response inferred from the system comparing to the response from familiar people. The naturalness can be rated by 5 grades, i.e., natural, a little natural, medium, a little unnatural, and unnatural. The confidence of ratting the naturalness is queried by 3 grades, i.e., confident, medium, and diffident. Part of the questionnaire is shown in Table 6.

Table 6. Part of the questionnaire

One day, the employee enters the secretary room. Employee*: Are you available now?* Secretary-robot*: Are you going to reserve a meeting-room?* Based on the background knowledge, The secretary-robot estimates that the employee is intend to reserve a meeting-room. How do you think of the response of this secretary-robot.
(1) natrual, (2) a little natural, (3) medium, (4) a little unnatural, (5) unnatural
How confident do you answering previous question? (1) confident, (2) medium, (3) diffident

Twelve subjects are invited to rate the eight output utterance from the five options, where "natural", "a little natural", "medium", "a little unnatural", and "unnatural" are assigned as 1, 0.75, 0.5, 0.25, and 0 respectively. The confidence are taken as weight of naturalness, where "confident", "medium", and "diffident" are mapped to 1, 0.5, and 0 respectively.

Table 7. Average rating of each output

Output utterance	Weighted average of naturalness
O1	0.84
O2	0.83
O3	0.74
O4	0.96
O5	0.65
O6	0.83
O7	0.92
O8	0.96
Average (AVG)	0.84

Table 8. Average rating of each subject

Subjects	Weighted average of naturalness
S1	0.81
S2	0.92
S3	0.91
S4	0.84
S5	0.91
S6	0.83
S7	0.77
S8	0.84
S9	0.66
S10	0.91
S11	1.00
S12	0.70
Average (AVG)	0.84

4.2 Results of Questionnaire Evaluation

The average result of each questions is shown in Table 7 where O1, O2...O8 mean outputted utterances marked in Tables 4 and 5. As shown in the Table 7, most of the output are evaluated between "a little natural" and "natural". Only two output (O3 and O5) are between "a little natural" and "medium". By applying the thoughtfulness and customized knowledge, proposal finally achieves a naturalness value of 0.84 which is between the ranks of "natural (=1.0)" and "a little nature (=0.75)" comparing to the response from familiar people in the same situation. Because of personal differences, some output utterances are rated lower than others. Since this employee goes to the secretary room for reserving meeting-room by a probability of 0.9, it could almost definite that the employee is about to reserve a meeting room when he enters the secretary-room and asks "Are you available". But some subjects still think that it will be natural to respond by "Can I help you", instead of "Are you going to reserve a meeting-room". When the employee looks sad and abnormal, some think that it will be more natural to ask "What's up" than "Please".

The evaluating result of each subject is shown in Table 8 where S1, S2...S12 stand for the twelve subjects. As shown in Table 8, ten of the twelve subjects rate the naturalness of utterance between levels of "natural" and "a little natural" while the rest of subjects just rate it as lower than the level of "a little natural".

It is concluded that the proposed deep level situation understanding may help to accomplish human-level natural communication in casual HRI.

The interaction between the human employee and the secretary robot in Tables 4 and 5 is also being demonstrated by DVD video [11] as shown in Figure 7. The perceived naturalness by the secretary robot is confirmed by the comparison experiment between the surface level emotion understanding by using voice, facial expression, and gesture of the human employee and the deep level emotion understanding with the customized knowledge.

5. Conclusions

A concept of deep level situation understanding is proposed for casual communications among humans and robots/machines. Twelve subjects are asked by questionnaire to evaluate the naturalness of the response of the proposed inference system comparing to the responses from familiar persons. The proposed system achieves a naturalness value of 0.84 which is in between the ranks of "natural (=1.0)" and "a little natural (=0.75)" comparing to communicate with familiar persons. It is concluded that the proposed deep level situation understanding may help to accomplish human-level natural communication in casual HRI.

Not only surface level understanding (e.g., speech/voice recognition, gesture/posture recognition), emotion understanding, intention understanding, and atmosphere understanding but also customized agent-dependent and universal knowledge, and a thoughtfulness mechanism are considered for smoothing and naturalizing communication among humans and robots/machines. The proposal can be applied to the service robot systems to achieve casual communication when interact with robots/machines. By considering the customized agent-dependent knowledge in human-robot communication, it will help robots/machines to understand the normal way in commu-

(a)

(b)

(c)

Figure 7. Interaction between the employee and the secretary-robot.

nication and to avoid unnecessary troubles and misunderstandings. With the comprehensive consideration of speech/voice, gesture/posture, emotion, intention, atmosphere, and knowledge (including universal and customized knowledge), the proposal will smooth the communication among humans and robots/machines as well as create a peaceful, pleasant, and prosperous society consisting of humans and various specification robots.

More and more, robots are required to do house work, care the elder, look after children, and work in the office. Hence the ability to communicate with persons at all ages is becoming essential. The proposal can smooth the communication among humans and robots by considering necessary knowledge (e.g., thoughtfulness and customized agent-dependent knowledge) for avoiding misunderstanding. Furthermore the proposed deep level situation understanding may help to build the coexistence and co-prosperity in the human-robot society in the near future.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This research project is supported by Japan Society for the Promotion of Science (No. 21300080).

References

[1] M. Sasajima, T. Yano, and Y. Kono, “EUROPA: a generic framework for developing spoken dialogue systems,” in *Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech1999)*, Budapest, Hungary, 1999, pp. 1-4.

[2] J. D. Williams, I. Arizmendi, and A. Conkie, “Demonstration of AT&T “Let’s Go”: a production-grade statistical spoken dialog system,” in *Proceedings of 2010 IEEE Spoken Language Technology Workshop (SLT)*, Berkeley, CA, 2010, pp. 157-158. http://dx.doi.org/10.1109/SLT.2010.5700839

[3] C. Shan, T. Tan, and Y. Wei, “Real-time hand tracking using a mean shift embedded particle filter,” *Pattern Recognition*, vol. 40, no. 7, pp. 1958-1970, 2007. http://dx.doi.org/10.1016/j.patcog.2006.12.012

[4] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, “Social signal processing: state-of-the-art and future perspectives of an emerging domain,” in *Proceedings of the 16th ACM International Conference on Multimedia*, 2008, pp. 1061-1070. http://dx.doi.org/10.1145/1459359.1459573

[5] C. Keskin and L. Akarun, “STARS: sign tracking and recognition system using input-output HMMs,” *Pattern Recognition Letters*, vol. 30, no. 12, pp. 1086-1095, 2009. http://dx.doi.org/10.1016/j.patrec.2009.03.016

[6] Y. Tang, H. A. Vu, P. Q. Le, D. Masano, O. Thet, C. Fatichah, Z. Liu, M. Yamaguchi, M. L. Tangel, F. Dong, Y. Yamazaki, and K. Hirota, "Multimodal gesture recognition for mascot robot system based on choquet integral using camera and 3D accelerometers fusion," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 15, no. 5, pp. 563-572, 2011.

[7] S. Hommel and U. Handmann, "AAM based continuous facial expression recognition for face image sequences," in *Proceedings of 2011 IEEE 12th International Symposium on Computational Intelligence and Informatics (CINTI)*, Budapest, Hungary, 2011, pp. 189-194. http://dx.doi.org/10.1109/CINTI.2011.6108497

[8] K. Shimada, K. Iwashita, and T. Endo, "A case study of comparison of several methods for corpus-based speech intention identification," in *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics(PACLING2007)*, Melbourne, Australia, 2007, pp. 255-262.

[9] Z. T. Liu, M. Wu, D. Y. Li, L. F. Chen, F. Y. Dong, Y. Yamazaki, K. Hirota, "Concept of fuzzy atmosfield for representing communication atmosphere and its application to humans-robots interaction," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 17, no. 1, pp. 3-17, 2013.

[10] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," in *Proceedings of Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference (APSIPA ASC 2009)*, Sapporo, Japan, 2009, pp. 131-137.

[11] J. A. G. Sanchez, K. Ohnishi, A. Shibata, F. Dong, and K. Hirota, "Deep level emotion understanding using customized knowledge for human-robot communication," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 19, no. 1, pp. 91-99, 2015.

[12] K. Ohnishi, F. Dong, and K. Hirota, "Atmosphere understanding for humans robots interaction based on SVR and fuzzy set," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 18, no. 1, pp. 62-70, 2014.

[13] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain, 2004, pp. 230-237.

Yongkang Tang received the Ph.D. degree from department of computational intelligence and systems science, Tokyo Institute of Technology (Tokyo Tech) in 2013. Currently he is a technical staff the department of Applied Informatics, Hosei University, Japan. His research interests include computational intelligence, robotics, and human robot interaction.
E-mail: tangyk@gmail.com

Fangyan Dong received Dr.E. degree from department of computational intelligence and systems science, Tokyo Institute of Technology (Tokyo Tech), Japan, in 2003. Since 2003, she has been with Tokyo Institute of Technology as a post-fellow researcher, an assistant professor, and currently is an associate professor of both ACLS (education Academy of Computational Life Sciences) and department of computational intelligence and systems science at Tokyo Institute of Technology. Her research interests include computational intelligence, logistics optimization, Kansei engineering, and intelligent robot. She is members of Japan Society for Fuzzy Theory and Intelligent Informatics, Japanese Society for Artificial Intelligence, and Information Processing Society of Japan. She published 75 journal papers and 130 conference papers, and received 10 awards.
E-mail: tou@acls.titehch.ac.jp

Yoichi Yamazaki received B.E. degree in Mechanics Engineering from TUS, the Tokyo University of Science in 2004. At TUS he was a student at Kobayashi laboratory. He had been enrolled in Tokyo Tech, the Tokyo Institute of Technology, where he has gotten a Dr.Eng. degree at Hirota laboratory. In 2009, He joined Kanto Gakuin University, and joined Kanagawa Institute of Technology in 2013. Currently, he is Associate Professor of Department of Home Electronics at Kanagawa Institute of Technology, where he is mainly involved in the research fields of communication robotics, home electornic system, mentality expression, and human-robot iteraction.
E-mail: yamazaki@he.kanagawa-it.ac.jp

Takanori Shibata received B.S., M.S., and Ph.D. in Electronic and Mechanical Engineering from Nagoya University in '89, '91, '92, respectively. He was a research scientist at AIST '93 to '98. Concurrently, he was a research scientist at the Artificial Intelligence Lab., MIT from '95 to '98. He has been a senior research scientist at AIST since '98. He was a Deputy Director for Information and Communication Technology Policy, Cabinet Office, Government of Japan from 2009 to 2010. His research interests include human-robot interaction, robot therapy, and humanitarian de-mining.
E-mail: shibata-takanori@aist.go.jp

Kaoru Hirota received Dr.E. degree from Tokyo Institute of Technology in 1979. After his career at Sagami Institute of Technology and Hosei University, he has been with Tokyo Institute of Technology. His research interests include fuzzy systems, intelligent robot, and image understanding. He experienced president and fellow of IFSA (International Fuzzy Systems Association), and president of SOFT (Japan Society for Fuzzy Theory and Intelligent Informatics.) He is a chief editor of J. of Advanced Computational Intelligence and Intelligent Informatics. Banki Donat Medal, Henri Coanda Medal, Grigore MOISIL Award, SOFT best paper award, Acoustical Society of Japan best paper award, honorary/adjunct professorships from "de La Salle University (Philippine), Changchun Univ. of Science & Technology (China), Harbin University of Science and Technology (China), the University of Nottingham (UK), and Beijing Institute of Technology (China)", and Honoris Causa from "Bulacan state university (Philippine), Budapest Technical University (Hungary), and Szechenyi Istvan University (Hungary)" were awarded to him. He organized more than 10 international conferences/symposiums as a founding/general/program chair. He has been publishing more than 250 journal papers, 50 books, and 500 conference papers.
E-mail: hirota@hrt.dis.titech.ac.jp