Human-Agent Groups: Studies, Algorithms and Challenges:
AAAI Technical Report FS-17-04

Proactive Conversation between Multiple Robots to Improve the Sense of Human–Robot Conversation

Yuichiro Yoshikawa,[1,2] Takamasa Iio,[1,2] Tsunehiro Arimoto,[1,2]

Hiroaki Sugiyama,[3] Hiroshi Ishiguro[1,2]

1) Dept. Systems Innovation, Graduate School of Engineering, Osaka University
2) ISHIGURO Symbiotic Human-Robot Interaction Project, JST, ERATO
3) Communication Science Laboratory, NTT
{yoshikawa, iio, arimoto.tsunehiro, ishiguro} @irl.sys.es.osaka-u.ac.jp
sugiyama.hiroaki@lab.ntt.co.jp

Abstract

In this position paper, we address potential merits of a novel conversational system using the group form of multiple robots that provides users with a stronger sense of conversation, with which a person can feel as if he or she is participating in a conversation. The merits can be performed by implementing the group behavior of multiple robots so that appropriate turn-taking is inserted to enhance the sense of conversation against potential conversational break-down. Through introducing the preliminary analysis of three experiments, how the sense of conversation can be enhanced and evaluated is exemplified and its limitations and potentials are argued.

Introduction

In recent years, conversational robots have received broad attention as a novel and effective media for education (Takana et al. 2014; Nakamura et al. 2015), entertainment (Softbank), and elderly care (Kuwamura et al. 2016; Kanoh et al. 2011). It is a formidable challenge for such robots to provide users with sufficient sense of a conversation, or a feeling of participating in a conversation.

Thus, robots must ensure that users do not consider robots as unable to logically grasp or disrupt a conversation. It is well known that robots can use nonverbal signals, such as nodding (Watanabe et al. 20103) or ambiguous replies (Weizenbaum 1966; Matsusaka et al. 2001), to provide humans with a feeling of being attended to instead of verbal sentences logically reflecting the context. However, such an approach of using ambiguous

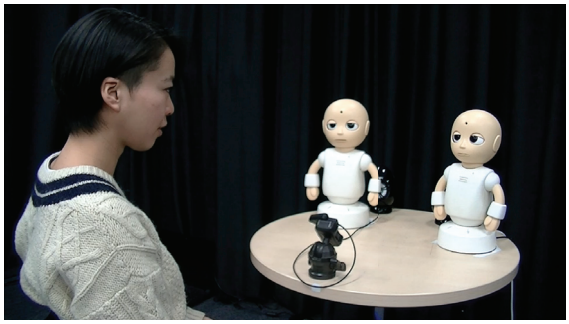

Figure1: Multiple conversational system

responses is not sufficient in encouraging humans to engage in a long-term conversation because of the lack of new information in the response. Meanwhile, studies have been conducted on the automatic mechanism of generating response sentences by finding an output sentence associated with sentences whose patterns match with the latest human utterances based on heuristic or statistical rules within human conversation (Higashinaka et al. 2014; Sugiyama et al. 2013). Although some of the previous methods can often generate successful sentences even in the open-domain conversation, it is inevitable to sometimes produce sentences that are difficult to interpret because there is no guarantee to find an appropriate pattern by focusing on the latest utterances. Furthermore, owing to the limitation in state-of-the-art artificial voice recognition, the recognized sentence of human utterance to be responded to includes errors. Therefore, it has not been easy for robots to perfectly avoid generating utterances whose relevance to the context is difficult to be interpreted. In other words, a conflict exists between necessity to consider human utterance and risk to generate inappropriate response; this

easily reduces human expectation or motivation toward robot conversation.

In this study, we focused on the potential merits of multiparty conversation enabled by adopting multiple robots instead of limiting ourselves to use a single robot as an interlocutor of humans. In this study, we addressed three dialogue patterns influential for the maintenance of conversation in a small-party conversation; these contributed toward preventing the occurrence of conversational breakdown. In this paper, we introduce a method of implementing the conversation with multiple robots. Moreover, this position paper introduces the experiments with the preliminary analysis without referring to these details in order to provide readers with a clear image about the effect of multiple robots on what we call sense of conversation, with which a person can feel as if he or she is participating in a conversation. Finally, the general discussion is given.

Automatic Structure against Breakdown Embedded in a Small-Party Conversation

The conversation ("Taiwa" in Japanese) is regarded as the process of exchanging concepts of values or finding their similarities between communication partners (Hirata 2012). Therefore, for a person to have the sense of conversation while talking with a robot, the robot must be able to know the types of value concepts the person intends to focus on and thus represent its intention to be shared. However, it is not easy to recognize an utterance used as the basic cue to recognize human intention. In other words, it is inevitable for a robot to produce irrelevant utterances such that humans fail to imagine the intention of the robot at all. Therefore, it is challenging to build a conversational system that can motivate humans to talk to it without any conversational breakdowns. However, in this study, we suppose that the conversational breakdown is inevitable but relaxed. Here, we address the three inherent structures within the small-party conversation (see Fig. 1); these relax the negative influence of the utterance with less relevance on the sense of conversation.

Bias in accepting conversation between others

In a small-party conversation consisting of more than two persons, a participant can be satisfied in participating in the conversation and obtaining information only by representing agreement/disagreement to the conversational topic instead of the explicit opinion. It is considered that even if humans participate in a conversation in such a way, they can imagine the sense of conversation in the conversation lead by others. This consideration has been utilized for the information technology using multirobot conversation. Sentences exchanged between two agents are known to be more understandable (Kubota et al. 2002), acceptable (Suzuki et al. 2004), and attractive (Sakamoto et al. 2009) than the corresponding ones read out by a single agent. A scene in which a robot utters a response to that uttered by another robot (see Fig. 1(a)) shows technical merit because it does not involve the process of recognizing human voice.

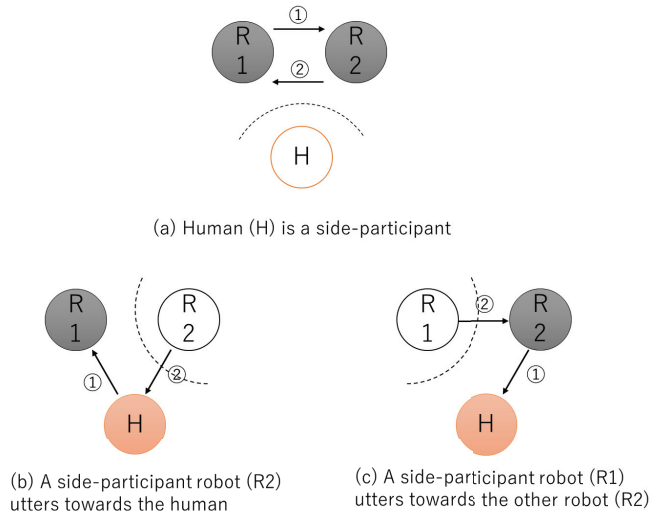

Figure 2: three structures inherent in small-party conversation

Bias regarding relevance in the utterance of the side-participant

In the small-party conversation, a participant has a choice to be a side-participant, that is, a participant spending time to not generate any positive utterances. In other words, being a side-participant has the advantage of a time margin, within which he or she can think of a current topic or topics further associated with the current topic. Therefore, being a side-participant might induce the person to be more influential for the next topic change. As participants holding the conversation have spent their cognitive resources for the past development of conversation, it is not usually easy for them to accurately estimate what the side-participant has contemplated. Therefore, when they receive the interruption from the side-participant, they cannot reject the possibility its reasonable relevance even when it seems to be causing a significant change in topic. As the result, the utterance of the side-participant is recognized as more relevant than the actual utterance. This is what we call as the bias regarding relevance in the utterance of the side-participant.

When a conversational system attempts responding to a human utterance, compelling a side-participant robot to utter a response is considered to be easier for relaxing the potential negative influence of the contextual gap on the conversation sense. Such an event is practically described as follows (see Figure 2(b)). i) Human H responds to robot R1 and ii) another robot used as a side-participant assumed

the speaker role. In this case, even though the utterance of R2 has less relevance to the last utterance of human H, it is unlikely that human H regards the utterance as irrelevant and the conversation is broken down.

Bias regarding relevance following up another

It is sometimes difficult for us to keep uttering a response if the utterance of other persons is not agreeable or attractive. In the small-party conversation, even in such an uncomfortable case, participants can easily confine their utterance by giving the speaking right to others until they are faced with an agreeable topic. In other words, in the small-party conversation, all participants do not have to accept all the narratives developed. This implies that when the robot utterance includes less relevance and human participant considers not accepting it, allowing the other robot to utter in a responsive way to the first robot utterance is reasonable.

This situation is practically described as follows (see Figure 2(c)). i) Robot R2 utters a response to a human utterance and ii) side-participant robot R1 produces an utterance in response to R2 in a way of scheduled harmony. Here, the utterance of the other robot produced immediately after one of the first robot has a side-effect of preventing humans from showing the irreparable assertion to discard opportunities to keep or return to participating to the conversation. Furthermore, even when the first robot's utterance did not seem to be relevant at the onset timing, it might gain relevance after the second robot's utterance. This is what we call bias regarding relevance following up another.

Examples of proactive conversation

In this section, we introduce our three previous studies based on the above-mentioned inherent structures in the small-party conversation. To construct the small-party conversation with a strong conversation sense, we developed a humanoid robot called CommU under the collaboration with Osaka University and Vstone Co. Ltd. CommU is a small child-like, desktop robot, which possesses 14 DOFs (3 for the eyeballs, 3 for the neck, 1 for the mouth, 1 for the eyelids, 2 for each arm, and 2 for the waist). Its height is approximately 30 cm, and it has a speaker inside its chest. CommU can produce its voice by playing a synthesized child-like voice while moving its mouth synchronously. The small size of its body allows the designer to consider its application in daily life space without intimidating users.

To utilize the structures inherent in the small party conversation including multiple robots, the phase of the robots talking to each other and the phase of either talking a human participant should be clearly distinguished by humans. In other words, they should prompt humans not to interrupt when they are conversing while naturally promoting the humans to reply. Thus, CommU was carefully designed so that human participants can easily distinguish to whom a CommU is talking to during a conversation.

Namely, CommU is given rich DOFs, such as eyeballs, neck, and waist, for representing the direction of its gaze, as well as DOFs for representing the state of utterance, such as the mouth.

The dialogue system for multiple robots

In the conversations with the proposed system, the multiple robots are controlled to proactively respond to user's answer to robot's question so as to avoid the break-down of the conversations and improve the sense of the conversations. To achieve it, the robots have to be controlled to synchronize with each other while the utterances, movements and gazes of each robot should be nicely coordinated. Therefore, we developed a description method to determine the coordination within and between robots as well as a dialog system that produces behaviors of robots as represented in the written script in the description method.

Table 1 shows a part of the script which describe behaviors of two robots R1 and R2 to greet a human H1. The script includes three types of commands: First, /say, /gesture and /look is a behavior command to produce an utterance, a gesture and a gaze, respectively. Second, /wait, /set and /sync are management commands to adjust the onset timings of executing behavior commands. /wait is used to postpone the execution of subsequent behavior commands for the specified time. /set is a command to register the timing of the command specified just before it into an arbitrary label, and /sync is used to synchronize the onset of the command specified after it with the corresponding timing specified by the label. Third, /WAIT and /JOIN, which are used in the head of line, are commands to totally manage timings of robot's behavior. /WAIT pauses to execute the next line of the /WAIT for the specified time. /JOIN waits to execute the next line of the /JOIN until the all commands from the previous /JOIN to the current /JOIN are completed. The head of each line of the script,

except for /WAIT and /JOIN, means the robot ID which indicates the controlled robot.

For example, in the table (a), the first line means to make the robot R1 say "Hello". The command '/set T0' in the same line means to register that timing as the name 'T0'. The command '/sync T0' means to synchronize the commands described in the current line with the timing of 'T0'. Particularly, in the second line, the robot R1 starts to execute a bow gesture at the timing when the speech "Hello" has finished. In a similar way, the third line means that the robot R2 starts to say "Hello" after one second of the timing T0. The command /JOIN in the fifth line means to wait until the executions of the commands from the first line to the fourth line are completed. In the sixth line, after the completion of the executions, the system pauses for two seconds. After two seconds, the robot R1 starts to say "We are CommU."

Table 1. A script example.

Line	Command
1	R1 /say { Hello } /set T0
2	R1 /sync T0 /gesture bow
3	R2 /sync T0 /wait 1.0 /say { Hello } /set T1
4	R2 /sync T1 /gesture bow
5	/JOIN
6	/WAIT 2.0
7	R1 /say { We are CommU } /set T2
8	R2 /sync T2 /wait 1.0 /say { What's your name? }
9	/JOIN

Switching speaker roles to conceal the contextual gap

We conducted an experiment to evaluate the influence of the multiplication of conversational robots on the conversation sense. Although we are analyzing the experiment in the detail levels and have not yet published it, the preliminary part is introduced in this subsection. Practically, we used one or two desktop humanoid robots and prepared two conditions: (1) a *Single* condition in which a participant spoke with a single robot and (2) *Multiple* condition in which the participant conversed with two robots. In each condition, the participants experienced the conversation without voice recognition based on the predetermined scripts. To control the information contents from robot(s), we used the same texts for robot's utterances between conditions. In addition, we carefully designed a script of utterances by using ambiguous words so as not to violate the logical consistency against the user's various utterances. Therefore, it was assumed that a participant might feel that the robot(s) sometimes does not follow his or her intention. In this experiment, we evaluated the participant's impression about the sense of conversation disruption in terms of an impression that robot(s) ignored the user and one that the conversation was difficult.

Practically, when the human was to say something in the script, the robots could wait until he or she finished. The end timing of the human utterance was identified when the breath group was detected or when a certain time had passed after the robot uttered toward the human. The robot utterances consisted of pairs of question toward the participants and response to the participants' replies. The replies were ambiguous words, such as "Uh-huh," or self-answers. An example of the question and ambiguous word are "what kinds of food do you like?" and "I see", respectively. The gaze of the robot was controlled to direct toward the current speaker (the participant or other robot) of the conversation.

We examined the evaluations collected from 26 university students (M=19.4, SD=0.95 [y]) participating in the conversation with a robot or robots for approximately 5 min. It was found that the participants had less moments of sensing difficulty in establishing a conversation in the *Multiple* condition than in the *Single* condition. Further, it was determined that the participants scarcely felt that the robot(s) were ignoring them in the *Multiple* condition than in the *Single* condition. From this result, we suggest that the conversation with multiple robots has a possibility for improvement. Moreover, we also evaluated an impression, which was assumed to be associated with the improvement: a positive sense to continue the conversation and a sense that robot(s) have intentions.

Collaborative response to conceal errors in recognizing human voice or intention

In the above study, the robots used ambiguous words in their responses to a user. They are expected to reduce a risk producing inconsistent responses. However, to give a user strong sense of conversation, it is important that the robots produce explicit utterance such as one representing agreement or empathy with a user; for example, "I think so too, because bra-bra." To produce such a response, the robots need to correctly recognize user's voice or intention, but such recognitions sometimes fail and may cause the breakdown of a conversation. Is it possible to express empathy only when the recognitions succeeded and to retain a conversation consistently even if the recognitions failed? In this subsection, we introduce a representative part of our proposed method for this problem and experimental result although we have analyzed it in detail and published in an international conference (Iio et al. 2016).

We have proposed a solution to this problem, that is generating a social context in which a user is guided to find coherency of the robot's utterances, even though its response is produced according to incorrect recognition of user's speech. To generate such a social context, we de-

signed a novel turn-taking pattern, in which two robots behave according to a prescheduled scenario. We designed a collaborative response by two robots to conceal errors in recognizing user's voice or intention. The collaborative response is composed of the following steps:

1. **[Question]** A robot asks a user something.
2. **[Answer]** The user answer to the question.
3. **[Back-channel]** The robot gives back-channel feedback; for example, "I see" and "Uh-huh".
4. **[Interjection]** The robot says an interjection to itself; for example, "Um..." and "Hmm..."
5. **[Response]** Another robot responds to user's answer.

Figure 5 shows two cases of the collaborative response. They are the same except for user's utterance. The left case shows that the robots correctly recognized user's intention that he or she is positive. On the other hand, the right case shows that the robots failed to recognize user's intention; even though the user are negative, the robot recognized that he or she was positive. Interestingly, even though the robots failed the recognition in the right case, the conversation sounds quite coherent. This is because the robot which states an opinion seems to express empathy with another robot. On the contrary, the robot seems to have an empathy with a user if the recognition succeeded, while it seems to have an empathy with another robot if the recognition failed. Therefore, even if the recognition fails, a conversation is retained consistently.

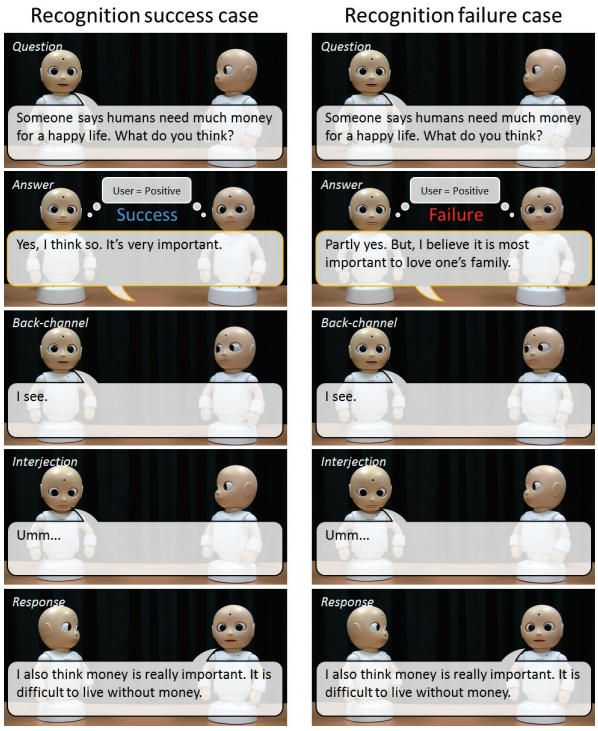

Figure 5: Two examples of a conversation using a collaborative response to conceal errors of the recognition.

We conducted an experiment to explore the efficacy of this collaborative response pattern we designed. The experiment was designed as within-participant design and had two conditions. One condition was one-robot condition, in which participants talked to a single robot that behave in a usual response pattern. The usual response pattern is almost same as the proposed collaborative response pattern except for not using interjections. The other condition was two-robots condition, in which participants talked to two robots that behave in the proposed collaborative response pattern. In this experiment, the robot(s) of both conditions said the same contents except for interjections according to a script we prepared for this experiment. 16 university students (M=21.5, SD=1.77 [y]) attended to the conversation in both conditions while the order of the conditions was counter-balanced. We had a human operator detect participant's voice activity instead of using an automatic detection technology (e.g. Wizard of Oz method). Just after the conversation in each condition, we asked four questions to evaluate the sense of conversation, using 7-point Likert scale to score the questions.

The question 1 on the felt coherence, "whether the participant felt robot's responses to his or her speeches almost made sense", was asked to evaluate a basic aspect of how successful the proposed method contribute to avoid losing the sense of conversation. We administrated Wilcoxon's sign rank test to check any differences in the scores for the above four questions between conditions. We found that participants had stronger feeling of coherence in two-robot condition than in one-robot one ($p < .05$) implying the positive sign of the effect of the proposed method.

Automatic insertion of short dialogue to prevent from losing joint attention to the focal point

We conducted an experiment on the system including the automatic mechanism to generate dialogue for multi-robot conversation. Although we are analyzing the experiment in the detail levels and have not yet published it, the preliminary part is introduced in this subsection. It is not generally easy for the conversational system to keep choosing appropriate replies. Thus, it is inevitable for human users to sometimes feel peculiar by the robot's response, and in the worst case, feel a conversational breakdown. Instead of merely speaker switching as described previously, we considered generating sentences for short dialogues to avoid losing joint attention to the focal point. In the proposed method, once the risk of an odd feeling is detected from the users toward the robot response, the system automatically generates new sentences from the original sentence that was to be given to users as the next utterance from either of the robots.

The sentence is padded or elongated without introducing additional information so as to allow users to observe a

short dialogue to recover the agreement of the focal point between the robots. (i) An ambiguous sentence is created by removing or replacing a part of the original sentence to demonstrative pronoun or simply removing and produced by a robot; (ii) the other robot asks the first robot to disambiguate the ambiguous sentence; (iii) the first robot disambiguates the sentence; and (iv) the second robot shows that it agrees to what the first robot is talking about. Some of the above-mentioned phases can be omitted. Such an insertion of the short dialog is expected to provide users with the feeling that the focal point of the conversation is maintained, as well as the opportunities of the second robot presumably having the same reaction as the human when the first robot utters something ambiguous, which will be resolved immediately through the inserted short dialogue. Furthermore, it can be effective for the designer to intentionally insert such sentences to enhance of feeling of joint attention on the focal point.

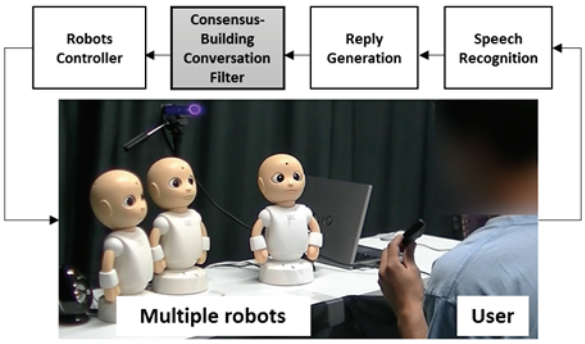

Figure 7: Conversational system with automatic padding insertion

We conducted a within-the-participant experiment to compare the impression of acceptability of the development of the conversational flow between two conditions in a small-party conversation using three robots: (a) *Original* dialogue condition and (b) *Padded* dialogue condition. In the *Original* condition, a popular chatter bot program, Docomo API dialogue (NTT Docomo Co., Ltd.) (NTT Docomo) was used to produce a reply to the participant's utterance. In the *Padded* condition, the same protocol was adopted, except that the short dialogue was inserted when the system judged whether the original sentence was extremely long, included the same focal word in the past, or a replaceable new focal word. Participants attended to the conversation of both conditions where the order of them were counter balanced. Just after each conversation, questionnaire with 7-point Likert scale about various subjective evaluation for the conversation. A question about the "difficulty to accept the topic changes" was asked to evaluate a basic aspect of how successful the proposed method contribute to avoid losing the sense of conversation. . The result obtained from the nine participants of university students (M=18.9, SD=1.05 [y]) whose score of social skill were ranked as lower than average revealed a statistical significant improvement in the subjective evaluation of difficulty to accept the topic changes in *Padded* condition than the *Original* condition ($p < .05$).

General discussion

The preliminary results in the experiments introduced in the previous section indicates the positive sign of the effect of using multiple robots on the sense of conversation. It is worth noting that it happens without adding any explicit information. It implies the effect is caused not by the contents but the way to convey it, which become available in the multi-robot conversation, that is the dialogue pattern including turn-taking. It would promise the potential merit for the service providers in implementing the information service since it allows them to focus on preparing the contents independently of how to convey it, which would be determined by tuning the dialog pattern of multiple robots. However, it is worth noting that the results introduced in this paper are limited in terms of cultural dependency because they were conducted in Japanese only for Japanese participants.

On the other hand, in this position paper, we could not sufficiently mention to the psychological phenomena behind the influence on the sense of conversation. However, various psychological or cognitive process such as empathy to the side-participant robot, cognitive dissonance, and post-diction are considered to take the important roles in this process. Therefore, computational modelling of it is the one the most important future work to correctly understand it as well as utilize it for more influential and successful conversational system through piling up the subjective experiments by controlling not only cultural but also the psychological backgrounds.

Furthermore, the voice activity detection took a crucial role for all three experiments. If the person speaks in an intermittent way, the robot or the operator might incorrectly detect the end of utterance of the user and interfere the user's utterance by starting its utterance too earlier. In the engineering viewpoint, extending the state-of-the-art for the method of verge-in to be more reliable and sophisticated one possibly by understanding the constraint in group dynamics is the one of the most important future work.

Conclusion

In this study, we analyzed dialogue patterns influential for the maintenance of conversation in a small-party conversation consisting of one person and more than one robot in

order to obtain ideas for building more conversational systems that relax the problem of conversational breakdowns. Namely, we addressed three possible constraints embedded in the small-party conversation: (i) bias in accepting a story among others, (ii) bias regarding relevance in the utterance of the side-participant, and (iii) bias regarding relevance following up others. In addition, three empirical studies were introduced to partially support these hypothetical constraints and exemplify the possibility of the approach to build more robust conversational systems in the real world. Note that the results introduced in this paper are limited in terms of cultural dependency because they were conducted in Japanese for Japanese participants. Further experiments should be conducted to make the model and conversational pattern more sophisticated and detailed.

Acknowledgments

This work was supported by JST ERATO Grant Number JPMJER1401, Japan. Part of this work was done under the collaborative research project among Osaka University, Communication Science Laboratory, NTT, and Advanced Telecommunication Research Institute International.

References

Tanaka, F.; Takahashi, T.; Tatsuzoe, S.; Tazawa, N.; and Morita, M. 2014. Telepresence robot helps children in communicating with teachers who speak a different language. In *Proceedings of 9th ACM/IEEE Intl. Conf. on Human-Robot Interaction*, 399-406.

Nakamura, S.; Watanabe, M.; Yoshikawa, Y.; Ogawa, K.; Ishiguro, H. 2015. Relieving Mental Stress of Speakers using a Tele-Operated Robot in Foreign Language Speech Education, In *Proceedings of Interspeech* 2015.

Softbank, http://www.softbank.jp/robot/consumer/products/

Kuwamura, K.; Nishio, S.; Sato, S. 2016. Can We Talk through a Robot As if Face-to-Face? Long-Term Fieldwork Using Teleoperated Robot for Seniors with Alzheimer's Disease. *Frontiers in Psychology* 7(1066): 1-13, July, 2016.

Kanoh, M., et al. 2011. Examination of practicability of communication robot-assisted activity program for elderly people. *Journal of Robotics and Mechatronics* 23(1): 3

Weizenbaum, J. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1): 36-45.

Matsusaka, Y; Fujie, S.; Kobayashi, T. 2001. Modeling of conversational strategy for the robot participating in the group conversation. In *Proceedings of Seventh European conference on speech communication and technology*.

Watanabe, T.; Danbara, R.; and Okubo, M. 2003. Effects of a Speech-Driven Embodied Interactive Actor "InterActor" on Talker's Speech Characteristics. In *Proc. of 12th IEEE International Workshop on Robot-Human Interactive Communication*

Higashinaka, R.; Imamura, K.; Meguro, T.; Miyazaki, C.; Kobayashi, N.; Sugiyama, H.; Matsuo, Y. 2014. Towards an open-domain conversational system fully based on natural language processing. In COLING: 928-939.

Sugiyama, H.; Meguro, T.; Higashinaka, R.; Minami, Y. (2013). Open-domain Utterance Generation for Conversational Dialogue Systems using Web-scale Dependency Structures. In *SIGDIAL Conference* (pp. 334-338).

O. Hirata 2012. Wakari Aenai Kotokara – What is the ability of communication -, *Gendai Shinsho Kodansya*

Kubota, H.; Yamashita, K.; Fukuhara, T.; Nishida, T. 2002. POC caster: Broadcasting agent using conversational representation for internet community. *Transactions of the Japanese Society for Artificial Intelligence*, 17: 313-321

Suzuki, S. V.; Yamada, S. 2004. Persuasion through overheard communication by life-like agents. In *Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology*: 225-231

Sakamoto, D. et al. 2009. Humanoid robots as a broadcasting communication medium in open public spaces. International *Journal of Social Robotics* 1(2): 157-169

Iio, T.; Yoshikawa, Y.; Ishiguro, H. 2016 Pre-scheduled turn-taking between robots to make conversation coherent. In *Proc. Fourth International Conference on Human Agent Interaction*: 19-25

NTT Docomo, Docomo API dialogue, https://www.nttdocomo.co.jp/service/developer/smart_phone/analysis/chat/