FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction
14-15 July 2018, Stockholm, Sweden

Comparison on Effect of Eye Gaze Activities between Human-human and Human-robot Conversations in Second-Language*

Koki Ijuin[1], Shohei Fujio[1], AlBara Khalifa[1], Tsuneo Kato[1], Seiichi Yamamoto[1]
[1] Graduate School of Science and Engineering, Doshisha University
1-3 Tatara, Miyakodani, Kyotanabe-shi, Kyoto 610-0394, Japan
euq1101@mail4.doshisha.ac.jp ctwb0107@mail4.doshisha.ac.jp albara.khalifa@gmail.com
tsukato@mail.doshisha.ac.jp seyamamo@mail.doshisha.ac.jp

Abstract

This paper examines how eye gaze activities are different in between human-human and human-robot conversations in second language (L2). The results show that the mainly-gazed-at listener gazes more at the speaker and he/she takes more often a floor in L2 conversations than in L1 conversations, whereas the speaker's eye gaze activity is almost the same in both conversations. The result shows that there is a significant positive correlation between the mainly-gazed-at listener's gazing ratio and the ratios of mainly-gazed-at listener taking a floor. Comparative analyses of eye gaze activities between human-human and human-robot conversations are also conducted. The results show that the listener gazes more at the speaker in human-robot conversations than in human-human conversations, whereas the robots do not provide the nonverbal information related to the contents of the utterances. These results may show that listeners gaze more at the speaker to show their intention to take a floor in both human-human and human-robot conversations.

1 Introduction

Gaze is one of the strongest and most extensively studied visual cues in face-to-face interaction among various nonverbal modality, and it is associated with a variety of functions, such as managing the attention of interlocutors [Vertegaal *et al.*, 2001], expressing intimacy and exercising social control, highlighting the information structure of the propositional content of speech, and coordinating turn-taking [Duncan, 1972], [Kendon, 1967].

These findings on human-human interactions were mainly obtained from conversations held in the native language (L1), and little is known of the effect of linguistic proficiency on multimodal conversations. The proficiency of conversational participants typically ranges widely from low to high in second-language (L2) conversations. As for eye gaze in L2 conversations, which is expected to have almost the same functionality as it has in L1 conversations, Hosoda [Hosoda, 2006] suggested that language expertise may affect the functions that eye gaze produces. Veinott et al. [Veinott *et al.*, 1999] found that non-native speaker pairs benefited from using video communication in route-guiding tasks, whereas native speaker pairs did not. These observations suggest that eye gaze may be more important in L2 conversations from those in L1 conversations.

Yamamoto et al. created a multimodal corpus of three-party conversations in L1 and L2 conversations conducted by the same interlocutors [Yamamoto *et al.*, 2015]. They showed that the averages of speaker's gazing ratios are almost the same in both conversations, whereas the averages of listener's gazing ratios are larger in L2 conversations than in L1 conversations. Ijuin et al. [Ijuin *et al.*, 2018] compared the speaker's gaze activities in L1 and L2 from the perspective of conversational interaction and suggested that the speaker's eye gaze tended to concentrate on one listener who is to be the next speaker, which resulted in an imbalanced amount of gaze between the two listeners, and the imbalance becomes larger in L2 conversations.

Eye gaze activities between human and robots are pointed out to play important role in human-robot interaction. Kozima and Ito showed that joint attention of human and robot by eye gazes played the dominant role in making coherent discourse for sharing the location to which the participants are paying attention [Kozima and Ito, 1997]. Previous research on human-robot communications suggested that joint attention occurred when the participants had interest in the gazed-at object [Ishii and Nakano, 2008]. These research suggest that eye gaze activities also play important roles for human-robot conversations.

These findings in human-robot conversations are also obtained in L1 conversations. There are few research on nonverbal behavior in human-robot L2 conversations. Opportunity of speaking in L2 is increasing by the rapid globalization. Computer assisted language learning (CALL) systems are under development by various research institutes and introduction of robots to CALL systems has been tried to supply more friendly circumstances to language learners. Khalifa et al. [Khalifa *et al.*, 2016] created multimodal corpus of human-robot L2 conversation system for learning English

*The authors would like to thank Dr. Ichiro Umata of KDDI Research, Inc. and Dr. Kristiina Jokinen of AIST, Japan for their suggestions and the various discussion we had with them. This research was supported in part by a grant from the Japan Society for the Promotion of Science (JSPS) (No. 15K02738).

10.21437/AI-MHRI.2018-5

through the conversations with two robots. They thought that the system needs to notice the tottering of conversations and to repair the conversation flow from conversational troubles caused with low linguistic proficiencies of participants, and the eye gaze activities of learners might be the one of useful factors for the system to estimate the learner's state.

In this paper, we compare the eye gaze activities in human-robot and human-human conversations to explore a possibility of using eye gaze activities for the purpose. This paper is structured as follows. We introduce the multimodal corpora we used in Section 2, and present our method and analysis results for eye gaze activities in Section 3. Then, we discuss our results in Section 4 and conclude with a summary in Section 5.

2 Multimodal Corpora of Human-human and Human-robot Conversations

A multimodal corpus created by Yamamoto et al. [Yamamoto *et al.*, 2015] was used to conduct correlation analyses between the listeners' eye gaze toward a speaker. The data of this multimodal corpus were collected from triad conversations in Japanese as the interlocutors' native language and in English as their second language. Three subjects participated in a conversational group, sitting in a triangular formation around a table. Three head-mounted eye trackers (NAC EMR-9) were used to record eye gaze.

The multimodal corpus included input from a total of 60 participants (23 females and 37 males: 20 groups). They were Japanese university students who had acquired Japanese as their L1 and had learned English as their L2. The corpus contains two conversations with different topics in each language. Each conversation was carried out in approximately six minutes. The total number of conversations is forty for each language. The multimodal corpus was manually annotated in terms of the time spans for utterances, backchannel, laughing, and eye movements.

A human-robot conversation corpus which was collected by Khalifa et al. [Khalifa *et al.*, 2016] was used to compare eye gaze activities between human-human and human-robot conversations. They created the prototype of English learning system through the conversations with two robots which one robot plays the role of a teacher, and the other plays an advanced peer learner as shown in Figure 1. The Wizard-of-OZ method was used to create this corpus so that the contents and timing of robots' utterances are controlled by the experimenter during the conversations. The natural recordings of native English speaker are used as utterances of robots. The conversational scenarios were constructed by repetition of the teacher robot's question and answers by a learner or the learner robot. When the teacher robot asks the questions to a participant, both robots gaze at the participant until the participant replies to the questions. If the participants could not answer the questions, the robot repeats the questions. The all participants are 25 Japanese university students (5 females and 20 males) who learn English as a second language. The corpus were manually annotated time spans of the utterances of participants and two robots, and participants' eye gazes toward each robot.

Figure 1: Experimental setup for human-robot conversations

3 Analyses

3.1 Methodology of Analyzing Eye Gazing During Utterances

We used the gazing ratio of the speaker, defined by Ijuin et al. [Ijuin *et al.*, 2018], for the classification of the listeners. Gazing ratios represent the ratios of how long the participant gazes at the other participant while the utterances. We classified the listeners into two groups according to their being targets of the speaker's eye gaze: mainly-gazed-at listener, who is gazed at more by the speaker during the utterance than the other listener, and "not-mainly-gazed-at listener." In the following sections, SPtoGL refers to the speaker's (SP) eye gaze toward the mainly-gazed-at listener (GL), while GLtoSP and NGLtoSP respectively refer to the mainly-gazed-at listener's and the not-mainly-gazed-at listener's (NGL) eye gazes toward the speaker.

3.2 Analyses of Speaker' Gazing Ratios towards Mainly-Gazed Listeners and Ratios of their Taking a Floor

To grasp the big picture of the participants' eye gaze activities, we calculated the average gazing ratios of SPtoGL, GLtoSP, and NGLtoSP in L1 and L2 conversations. As shown in Table 1, the gazing ratios of SPtoGL are almost the same in L1 and L2 conversations, whereas both gazing ratios of GLtoSP and NGLtoSP are higher in L2 conversations than in L1 conversations. To analyze the differences, we conducted an ANOVA test with language difference, topic difference and gaze channel (gazer-target pairs) difference being within-subject factors. The ANOVA test for SPtoGP does not show any significant main effect or interaction. The ANOVA test for GLtoSP and NGLtoSP shows significant main effects of language difference ($F_{(1,19)} = 52.5, p < .01$) and gaze channel difference ($F_{(1,19)} = 115.8, p < .01$). These results show that there is no difference in ratios of SPtoGL between L1 and L2 conversations, the ratios of GLtoSP is higher than ratios of NGLtoSP in both language conversations, and both listeners gaze more at a speaker in L2 conversations than in L1 conversations.

Table 1: Basic statistics of SPtoGL, GLtoSP, and NGLtoSP ratios in L1 and L2 conversations.

Gazing ratios	L1 conv.	L2 conv.
SPtoGL	52.5%	53.3%
GLtoSP	54.1%	66.0%
NGLtoSP	43.9%	54.2%

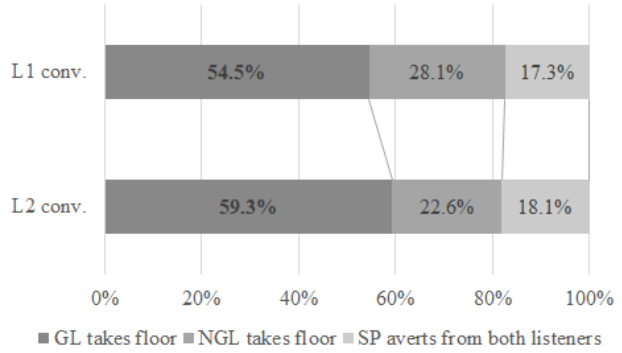

Figure 2: Ratios of which listener takes the floor in L1 and L2 conversations. Note that the utterances which the speaker did not gaze at any listener, that is, the utterances which listeners cannot be classified with the speaker's eye gaze , are classified into "SP averts from bot listener".

To verify the effects of speaker's eye gaze activities, we calculated the ratios of how often the mainly-gazed-at listener takes a floor. Figure 2 represents the ratios of which listener takes the floor in L1 and L2 conversations. The ANOVA on the ratios of mainly-gazed-at listener taking a floor was conducted with language difference and topic difference being within-subject factors. The results show significant main effect of language difference ($F_{(1,19)} = 7.2, p < .05$). The statistic results revealed that the mainly-gazed-at listener takes a floor more often in L2 conversations than in L1 conversations, whereas quantity that the speaker gazes at the mainly-gazed-at listener is almost the same in both L1 and L2 conversations.

The statistic results of human-human conversations suggest that the effect of speaker's eye gaze activities for floor apportionment is stronger in L2 conversations than in L1 conversations, and the mainly-gazed-at listener gazes more at the speaker and takes a floor more often in L2 conversations than in L1 conversations.

3.3 Comparison on Listener's Eye Gaze Activities between Human-human and Human-robot Interactions

Fujio et al. [Fujio *et al.*, 2018] compared the listener's eye gaze activities in utterances between human-human and human-robot conversations in the both corpora, and reported that the listener's gazing ratio is higher in human-robot conversations than in human-human conversations.

In this corpus of human-robot conversations, the teacher robot gazes only at the participant (learner) when questioning him/her, and the learner takes a floor after the utterances of questions.. To precisely compare eye gaze activities in human-human and human-robot conversations in similar condition, we calculated the gazing ratios of GLtoSP when the speaker gazes only at the mainly-gazed-at listener during his/her utterances in human-human conversations, and the gazing ratios of learner toward teacher robot during teacher robot's utterances in human-robot conversations. Figure 3 compares these gazing ratios. The paired-t test shows that there is a significant difference between human-human L2 conversations and human-robot conversations ($t_{(43)} = 3.9, p < .01$). The results shows that the mainly-gazed-at listener gazes more at speaker in human-robot conversations than in human-human L2 conversations.

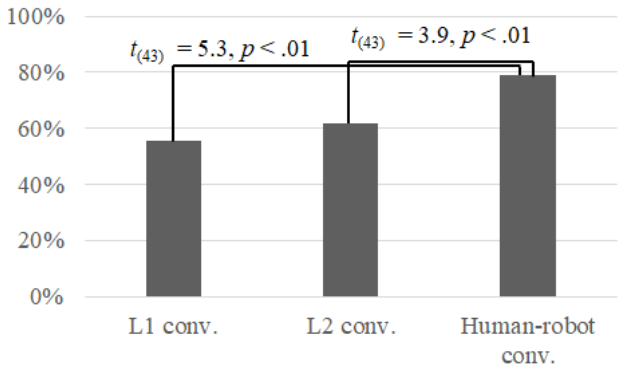

Figure 3: gazing ratios of the mainly-gazed-at listener toward the speaker in human-human L1, L2 conversations and human-robot conversations in case that the speaker gaze only at the mainly-gazed-at listener during his/her utterances.

4 Discussion

The analyses of eye gaze activities in human-human conversations demonstrated the mainly-gazed-at listener takes a floor more often in L2 conversations than in L1 conversations, whereas quantity that the speaker gazes at the mainly-gazed-at listener is almost the same in both L1 and L2 conversations. The result also showed that the listener who was gazed by the speaker gazed more at speaker than the other listener both in L1 and L2 conversations.

These results suggest that the speaker's eye gaze activity has strong effect in leading the mainly-gazed-at listener to the next speaker in L2 conversations, although the duration that the speaker gazes at the listener is almost the same in L1 and L2 conversations. This stronger effect of the speaker's eye gaze activities for determining the next speaker in L2 conversations may lead to the higher gazing ratio of the mainly-gazed-at listener toward the speaker in L2 conversations than in L1 conversations. The listeners may show their intention to take a floor by their eye gazes toward speaker, and when their linguistic proficiencies were lower, they gaze more at the speaker.

The comparative analyses of gazing ratios of mainly-gazed-at listeners between human-human and human-robot conversations demonstrated that the listener gazes more at the speaker in human-robot conversations than in human-human

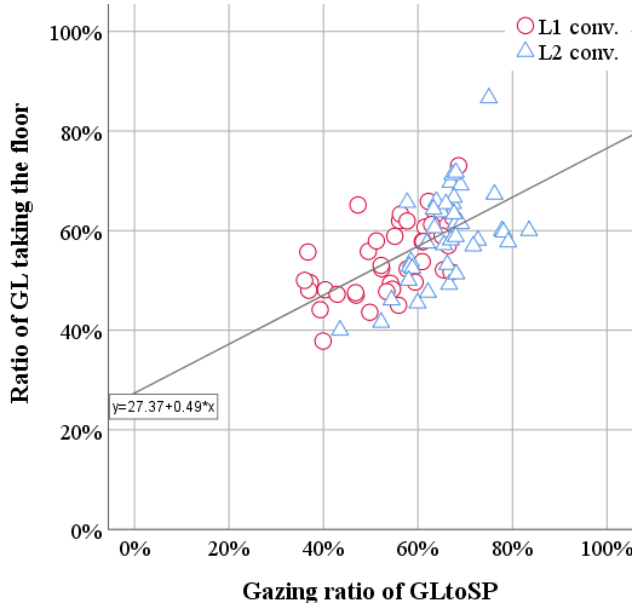

Figure 4: Scatter plots of gazing ratio of GL2SP and ratios of GL taking a floor in human-human L1 and L2 conversations.

conversations. Although the robots in this corpus do not provide any useful nonverbal information for understanding the contents of utterances, the participants still gaze at the speaker. There is a possibility that listeners gaze at the robot unconsciously to show their intention to take a floor even in human-robot conversations.

To confirm this possibility, we conducted the correlation analysis between gazing ratios of GLtoSP and ratios of the mainly-gazed-at listeners taking a floor, as shown in Figure 4. The result shows that there is a significant positive correlation between the gazing ratios of mainly-gazed-at listener and ratios of the mainly-gazed-at listeners taking a floor ($r = .62, p < .01$). This result suggests that the more mainly-gazed-at listeners gaze at the speaker, the more often they tend to take a floor in both language conversations. The human-robot conversations in this corpus were designed as scenario-based conversations so that the listeners were forced to take a floor after the robot's questions. This experimental set-up may increase the probability that the listeners show their intention unconsciously to take a floor with their eye gaze activities. This might be the one of the reason why the listeners in human-robot conversations in this corpus gaze more at a speaker than in human-human conversations.

5 Conclusion

We compare eye gaze activities between L1 and L2 conversations in human-human conversations. The results show that the mainly-gazed-at listener gazes more at the speaker and he/she takes more often a floor in L2 conversations than in L1 conversations, whereas the speaker's eye gaze activity is almost the same in both conversations. We compare eye gaze activities between human-human and human-robot conversations. The results show that the listener gazes more at the speaker in human-robot conversations than in human-human conversations, whereas the robots do not provide any nonverbal information related to the contents of the utterances. These results may show that the listener gazes more at the speaker unconsciously to show their intention to take a floor even in human-robot conversations.

References

[Duncan, 1972] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292, 1972.

[Fujio *et al.*, 2018] Shohei Fujio, Koki Ijuin, Tsuneo Kato, and Seiichi Yamamoto. Measurement of gaze activities of learners with joining-in-type rall system. In *Proceedings of the 2018 IEICE General Conference*, march 2018. (In Japanese).

[Hosoda, 2006] Y. Hosoda. Repair and relevance of differential language expertise in second language conversations. *Applied Linguistics*, pages 25–50, 2006.

[Ijuin *et al.*, 2018] K. Ijuin, I. Umata, T. Kato, and S. Yamamoto. Difference in eye gaze for floor apportionment in native- and second-language conversations. *Journal of Nonverbal Behavior*, 42(1):113–128, 2018.

[Ishii and Nakano, 2008] Ryo Ishii and Yukiko I. Nakano. *Estimating User's Conversational Engagement Based on Gaze Behaviors*, pages 200–207. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[Kendon, 1967] A. Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[Khalifa *et al.*, 2016] AlBara Khalifa, Tsuneo Kato, and Seiichi Yamamoto. Joining-in-type humanoid robot assisted language learning system. In *Proceedings of 10th International Conference on Language Resources and Evaluation*, may 2016.

[Kozima and Ito, 1997] Hideki Kozima and Akira Ito. The role of shared-attention in human-computer conversation. In *In International Conference of Research on Computational Linguistics (ROCLING-97)*, pages 224–228, 1997.

[Veinott *et al.*, 1999] E. Veinott, J. Olson, G. Olson, and X. Fu. Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. In *Proceedings of the Conference on Computer Human Interaction. CHI'99, ACM Press, PA, USA*, pages 302–309, 1999.

[Vertegaal *et al.*, 2001] R. Vertegaal, R. Slagter, G. Veer, and A. Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *CHI '01 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 301–308, 2001.

[Yamamoto *et al.*, 2015] Seiichi Yamamoto, Keiko Taguchi, Koki Ijuin, Ichiro Umata, and Masafumi Nishida. Multimodal corpus of multiparty conversations in l1 and l2 languages and findings obtained from it. *Language Resources and Evaluation*, 2015.