Multi-party Video Conferencing system with Gaze Cues Representation for Turn-taking

Rikuto Iitsuka[1], Ikkaku Kawaguchi[2], Buntarou Shizuki[2], and Shin Takahashi[2]

[1] University of Tsukuba, Tsukuba, Japan
iitsuka@iplab.cs.tsukuba.ac.jp
[2] University of Tsukuba, Tsukuba, Japan
{kawaguchi,shizuki,shin}@cs.tsukuba.ac.jp

Abstract. In a multi-party video conference, it is more difficult to achieve smooth turn-taking than in face-to-face communication. This is probably because gaze cues are not shared. In this paper, we propose a system for facilitating turn-taking through the sharing of gaze cues in multi-party video conferences. We implemented video conferencing systems that use arrows and modification of the video window size to share gaze cues the same as in face-to-face communication. We also conducted an experiment to investigate the effect of the system on turn-taking. The results suggested that our system could facilitate turn-taking and communication.

Keywords: multi-party video conference · remote communication · turn-taking · gaze interaction · participation status.

1 Introduction

In recent years, video conferencing systems such as Zoom [3] and Microsoft Teams [4] have been increasing. However, compared with face-to-face communication, it is difficult to achieve smooth turn-taking in a multi-party video conference, and speech contention and silence often occur. In comparison, in face-to-face communication, there is little speech contention and silence, and smooth turn-taking is achieved. The reason for this is that the participants in face-to-face communication know each other's participation status such as who the speaker is. Gaze cues (information on whom each participant is looking at) play an important role in understanding the participation status. However, in a multi-party video conference, gaze cues are not shared. Therefore, participants are unable to know each other's participation status, and that causes speech contention and silence in turn-taking.

To solve this problem, much research has been conducted to promote communication through the sharing of gaze cues in video conferences [5, 10]. However,

[3] https://zoom.us/

[4] https://www.microsoft.com/ja-jp/microsoft-365/microsoft-teams/free/

in these studies mainly real space or 3D CG space have been used to represent the relative positions and gaze directions of the participants, and little research has been conducted on video conferences with multi-party participants using 2D video windows such as Zoom and Microsoft Teams.

Therefore, in this paper, we propose a multi-party video conferencing system (Figure 1) that detects the gaze cues of each participant with a the display-based eye tracking program and visualizes them by using arrows or by modifying of the video window size.

The system aims to promote the understanding of participation status and is expected to reduce speech contention and silence during turn-taking in multi-party video conferencing.

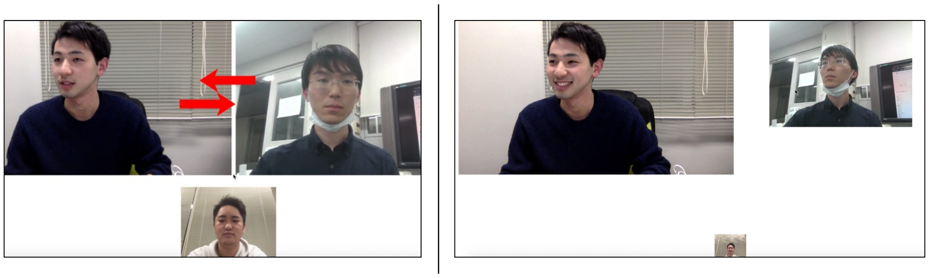

Fig. 1. Proposed multi-party video conferencing systems with two gaze cues representations. Shown are gaze cues represented by arrows (left) and by a change video window size (right). Left and right figures are representations of same gaze cues. Participant in upper left is gazing at participant in upper right, and participants in upper right and below are gazing at participant in upper left.

2 Related work

Nonverbal information plays an important role in face-to-face communication [9]. Kendon [4] showed that turn-taking is achieved when the speaker gazes at the addressee during a break in speech, and the addressee accepts the gaze and returns it to the speaker (mutual gaze). In addition, the gaze cues of the speaker before the end of an utterance indicate the intention to pass the floor and encourage listeners to be aware of their own participation status. Goffman [3] stated that *the participation status* that each participant has depends on the degree of participation in communication. Participation status include the *speaker* who is currently speaking, the *addressee* who should get the floor next, and the *side participant* who participates in the communication but does not become the addressee. The roles of participants in communication are determined by the gaze cues of the speaker. In addition, the listener's gaze cues influence the speaker's choice of the next speaker [11]. Therefore, participants understand their participation status on the basis of each other's gaze cues [1].

There are many works in which nonverbal information has been used to support turn-taking in video conferencing. Tamaki et al. [8] proposed a method that supports smooth turn-taking in video conferencing by detecting pre-motions before a person speaks and highlighting the participant who is most likely to speak next. A pre-motion is an action that a person performs before speaking, and it expresses the desire to speak. However, Tamaki et al. did not use gaze cues as the pre-motion. Vertegaal et al. [10] proposed a multi-party video conferencing system that transmits eye contact to participants. In this system, the video windows of each participant are arranged in 3D CG space on a the display, and the system turns the video windows to present gaze cues.

Okada et al. [6] conducted a video conference with three people by projecting an actual-size image of the other people on a curved screen and matching their gaze to it. Thus, much research has been conducted to promote communication by sharing gaze cues in video conferences. In these studies mainly real space or 3D CG space have been used to represent the relative positions and gaze directions of the participants. However, existing video conferencing systems such as Zoom and Microsoft Teams display video windows in 2D.

In this paper, we propose a multi-party video conferencing system in which the video windows of each participant are arranged in 2D. In addition, we propose a method for promoting the understanding of the participation status and for facilitating turn-taking in a multi-party video conference by transmitting gaze cues.

3 Proposed system

We implemented a multi-party video conferencing system with gaze cues representation as a Web application.

To implement the system, we used SkyWay [2], a WebRTC platform. SkyWay was also used to send and receive detected gaze directions. A JavaScript library, WebGazer.js [7], was used to detect gaze directions. WebGazer.js detects the point on the display at where the user is looking during a video conference with the built-in camera of a laptop. The detected gaze cues information is sent to each participant using SkyWay and expressed in accordance with each gaze cues representation.

In this research, the video conferencing system is for three participants. The system arranges each of their video window . The system was designed to make one's own window small and the windows of the other participants large. The reason for this was that the results of a preliminary experiment showed that the participants behaved differently from face-to-face communication (e.g., they gazed at their own video window) when the windows were arranged evenly.

Our implemented system represent participants' gaze cues and then shares gaze cues with other participants. We propose two gaze representation methods, which are described in section 3.2.

3.1 Gaze cues representation methods

We proposed two representation methods for visualizing gaze cues. The first represents gaze cues using arrows. The second represents gaze cues by changing the size of the video window.

Method for representing gaze cues with arrows Gaze cues of each participant are represented by an arrow (Figure 1, left). An arrow is displayed on the video screen when the participant is gazing at another participant. However, the arrow that shows who they themselves are gazing at is not displayed on their own screen. Since information on who is looking at whom is shared directly, gaze cues from the speaker can be clearly recognized.

Method for representing gaze cues by changing size of video window The video window that many participants are gazing at is enlarged, and the video windows of participants without gaze cues are reduced in size(Figure 1, right).

In this representation, the participant with the most gaze cues is displayed the largest, making it clear to the listener who to gaze at. There are three window sizes (when no one is gazing, one person is gazing, and two people are gazing), and it does not count when a participant gazes at their own video window. In other words, gaze cues are indirectly shared as the size of the video window. Also, the size of the video window indicates the participation status.

4 Experiment

Using the proposed system, we conducted an experiment to find out whether a video conference using it actually facilitated turn-taking. In this section, we describe the experimental design, results, and discussions.

4.1 Experimental design

Three participants held a discussion to generate as many ideas as possible by using the video conferencing system. A total of 9 participants (8 men and 1 woman, mean age of 21.7 years) participated. They were split into 3 groups of 3 participants. The following three conditions were set.

- **Control Condition** : video conferencing system with no gaze cues
- **Arrow Condition** : video conferencing system with gaze cues representation method using arrows
- **Window Condition** : video conferencing system with gaze cues representation method using the size of the video window

The experiment was conducted in a within-subject design. We used a Latin square method for counterbalancing and determined the order of conditions for each group. Each conference lasted for 7 minutes for each session. At the end of each session, participants completed a questionnaire using a 7-point Likert scale to canvass their subjective evaluation of whether the turn-taking and the conference were facilitated. Table 1 shows the items of the questionnaire.

Table 1. Questionnaire for investigating subjective evaluation of whether turn-taking and conference were facilitated.

	Items
A–1	I think the participants listened to my speech.
A–2	I think the participants found I listened to their speech well.
A–3	During a break in my speech, I found the next speaker.
A–4	During a break in another participant's speech, I found the next speaker.
A–5	During a break in another participant's speech, I could speak.
A–6	I found who is the speaker is well.

We recorded audio and video of the conferences, and we conducted a conversation analysis after the experiment. The ratio of failed turn-takings and the number of utterances of each participant were counted as items to evaluate whether the turn-taking and conference were facilitated. The ratio of the failed of the turn-takings was obtained by dividing the sum of the number of speech contentions and silences by the number of turn-takings.

4.2 Results

Conversation analysis First, we conducted a Shapiro-Wilk test and checked the normality (p >.05). We also conducted a Bartlett test and checked the homogeneity of variances (p >.05).

Second, we conducted a one-way ANOVA on the ratio of failed turn-takings among the three conditions (Figure 2, left), and there was a marginally significant effect of the condition (p <.10). We conducted a multiple comparison test with Bonferroni correction, and we found that the window condition was marginally significantly different compared with the control condition (p <.10).

Finally, we conducted a one-way ANOVA on the number of utterances of each participant among the conditions (Figure 2, right), and there was a significant effect($p = 0.0296$ <.05) of the condition. Under a corrected significance level, a multiple comparison test showed that the window condition was significantly different compared with the control condition ($p = 0.0442$ <.05). We also found the arrow condition was marginally significantly different than the control condition (p <.10).

Questionnaire for investigating subjective evaluation of the conference The results of the questionnaire are shown in Figure 3. These items were used

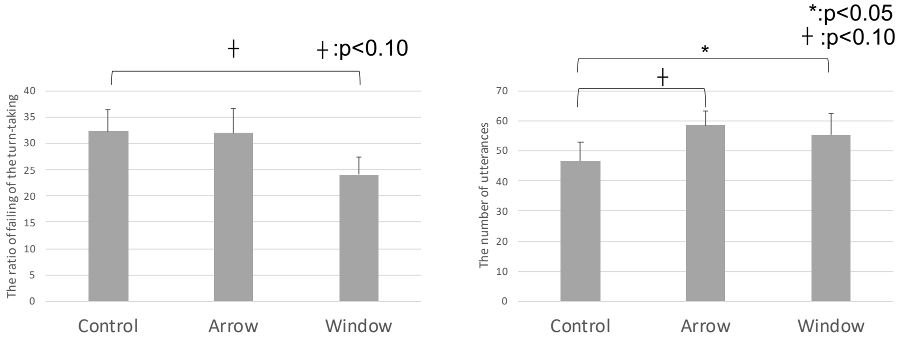

Fig. 2. Results of conversation analysis.Ratio of failed turn-takings (left) and the number of utterances of each participant among three conditions (right) are shown.

to subjectively evaluate whether the system was able to promote participants' understanding of their own participation status in the conference. The result of

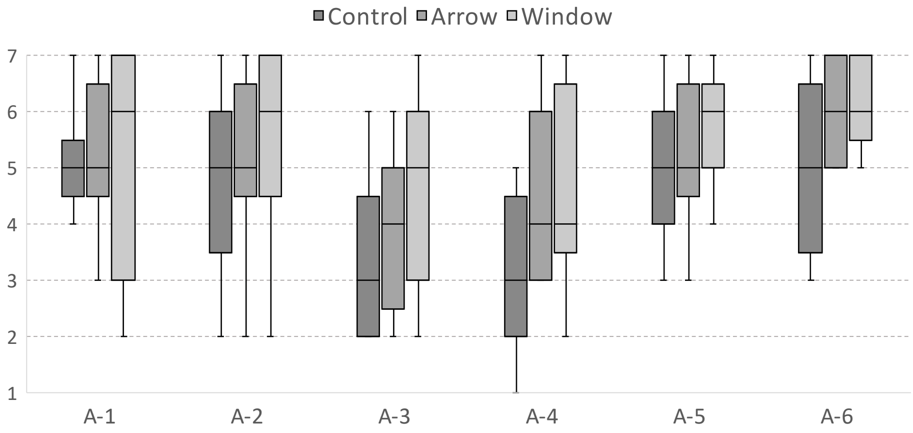

Fig. 3. The results of the questionnaire. These items were used to subjectively evaluate whether the system was able to promote the understanding of their own participation status in the conference. The results for A1 through A6 are shown.

a Friedman's test showed that there was no effect of the condition for all items. However, when the median values were compared, the window condition was rated higher than the control condition for all items. In the arrow condition, the median values of some items were the same as those in the control condition. While turn-taking was promoted, there was no effect of the condition for all items in the subjective evaluation. This could have been caused by a lack in the number of participants.

5 Discussion

In the arrow condition, turn-taking could not be promoted. We also found that participants often looked outside the display. The reason for this could be that the understanding of the participation status was not promoted. Therefore, it might be necessary to further promote this understanding of the participation status in the arrow condition; for example, it might be necessary to consider a design in which one's own gaze cue is displayed on one's video screen, or a design in which one's gaze direction is guided toward the display.

There was a possibility that the window condition promoted the understanding of the participation status. In addition, the number of failed turn-takings decreased, and the number of utterances increased, suggesting that the sharing of gaze cues promotes turn-taking and facilitates conferences in multi-party video conferences. The results of the window condition were also more significant than those of the arrow condition, suggesting that using this method to show the participation status may be effective.

In this experiment, 9 participants were split into 3 groups of 3 participants. However, there were some analysis results that showed no significant differences due to the a lack of data. Therefore, it is necessary to conduct an experiment with more groups as an additional survey to increase the number of pieces of data.

6 Conclusion

The purpose of this study was to reduce speech contention and silences and to facilitate turn-taking in a multi-party video conference. Therefore, we proposed a multi-party video conferencing systems that promotes the understanding of participant status through the sharing of gaze cues.

On the basis of the results of a preliminary experiment, we implemented a video conferencing system promoting gaze cues to be equivalent to those of face-to-face communication and sharing gaze cues. To share gaze cues, we implemented a gaze cues representation method using arrows and one using the size of the video window.

We conducted an experiment to find out whether the proposed system could promote the understanding of participation status and whether it facilitates turn-taking. As a result of the experiment, it was found that our system could facilitate turn-taking and the communication. We also discussed a subjective evaluation of each condition based on the results of a questionnaire, and our findings that will lead to improving the system in the future.

In the future, we plan to improve the gaze cues representation methods based on the findings and conduct experiments with a sufficient number of participants to verify the effectiveness of the proposed system in more detail.

References

1. Bono, M., Suzuki, N., Katagiri, Y.: An analysis of participation structure in conversation based on interaction corpus of ubiquitous sensor data. In: INTERACT. vol. 3, pp. 713–716 (2003)
2. Corporation, N.C.: Skyway. https://webrtc.ecl.ntt.com/
3. Goffman, E.: Replies and responses. Language in Society **5**(3), 257–313 (1976), http://www.jstor.org/stable/4166887
4. KENDON, A.: Some functions of gaze-direction in social interaction. Acta Psychologica **26**, 22–63 (1967). https://doi.org/10.1016/0001-6918(67)90005-4, https://ci.nii.ac.jp/naid/30008655637/
5. Mukawa, N., Oka, T., Arai, K., Yuasa, M.: What is connected by mutual gaze? user's behavior in video-mediated communication. In: CHI '05 Extended Abstracts on Human Factors in Computing Systems. pp. 1677–1680. CHI EA '05, Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1056808.1056995, https://doi.org/10.1145/1056808.1056995
6. Okada, K.I., Maeda, F., Ichikawaa, Y., Matsushita, Y.: Multiparty videoconferencing at virtual social distance: Majic design. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. pp. 385–393. CSCW '94, Association for Computing Machinery, New York, NY, USA (1994). https://doi.org/10.1145/192844.193054, https://doi.org/10.1145/192844.193054
7. Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., Hays, J.: Webgazer: Scalable webcam eye tracking using user interactions. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI). pp. 3839–3845. AAAI (2016)
8. Tamaki, H., Higashino, S., Kobayashi, M., Ihara, M.: Reducing speech contention in web conferences. In: 2011 IEEE/IPSJ International Symposium on Applications and the Internet. pp. 75–81 (2011). https://doi.org/10.1109/SAINT.2011.20
9. Vargas, M.F.: Louder than words : an introduction to nonverbal communication. Iowa State University Press (1986), https://ci.nii.ac.jp/ncid/BA0036671X
10. Vertegaal, R.: The gaze groupware system: Mediating joint attention in multiparty communication and collaboration. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 294–301. CHI '99, Association for Computing Machinery, New York, NY, USA (1999). https://doi.org/10.1145/302979.303065, https://doi.org/10.1145/302979.303065
11. Vertegaal, R., Slagter, R., van der Veer, G., Nijholt, A.: Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 301–308. CHI '01, Association for Computing Machinery, New York, NY, USA (2001). https://doi.org/10.1145/365024.365119, https://doi.org/10.1145/365024.365119