INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Analysis of eye gaze reasons and gaze aversions during three-party conversations

Carlos Toshinori Ishi [1,2], *Taiken Shintani* [1]

[1] RIKEN Guardian Robot Project, Japan
[2]ATR Hiroshi Ishiguro Labs, Japan
carlos.ishi@riken.jp, taiken.shintani@riken.jp

Abstract

The background of this study is the generation of natural gaze behaviors in human-robot multimodal interaction. For that purpose, in this study we analyzed gaze behaviors of multiple speakers in a dataset containing three-party conversations, in terms of the reasons/intentions of their gaze events.

Analyses of the gaze reasons were conducted separately for the gaze behaviors towards a dialogue partner, and for gaze aversions (i.e., gazing away from a person's face). Analysis on the eyeball movements during gaze aversions was also conducted. Different distributions for average durations and gaze direction patterns were observed depending on the gaze reasons (e.g., in listening mode, speaking mode, towards dialogue partner's reactions, in gaze aversions during thinking and remembering, and during the speaker's own behaviors like nodding and laughing).

Index Terms: eye gaze reasons, gaze aversion, spontaneous conversation, multimodal analysis, visual prosody

1. Introduction

Eye gaze has important functions in dialogue communication, such as to provide feedback, transmit emotional information, and regulate conversation flow (e.g., opening interactions and directing attention) [1, 2]. Therefore, it is important to suitably control eye gazing in dialogue agents and robots, in order to achieve smooth communication with humans.

Numerous studies have been conducted on eye gaze in human-agent and human-robot interactions so far (extensive reviews of social eye gaze can be found in [3, 4]). Social gaze can be roughly classified in mutual gaze (eye contact), referential or deictic gaze (towards an object or location), joint attention (to a common object) and gaze aversions (shifts of gaze away from the main direction of gaze, which is typically a partner's face) [4].

Regarding regulation of conversation flow, several studies have reported correlations between turn-taking and gaze behaviors in dyadic and multi-party dialogue interactions [5] [6] [7]. Other studies have shown that eye gaze along with other cues are important for estimating engagement [8, 9], and interpersonal reactivity scores (empathy skill levels) during turn-changing [10]. The conversational dominance is also considered as a factor affecting gaze behaviors, besides the dialogue roles (speaker, addressee, or side participant), in the models proposed in [11].

Although most of studies on gaze generation have considered the turn-taking functions in dialogue interactions, there are fewer studies explicitly modeling gaze aversions [6] [12]. Furthermore, there are fewer studies focusing on the meanings and reasons of the gaze movements, and clarifying how the gaze behaviors change according to different dialogue situations.

Thus, in this study, we conducted analysis of the reasons/intentions behind the gaze behaviors in three-party dialogue interactions, separately for gazing directed to a dialogue participant's face, and gaze aversions (i.e., gazing not directed to a person's face). We also conduct detailed analysis on the eyeball movements during gaze aversions.

2. Analysis data

For analysis of gaze behaviors, we used a dataset of a multimodal three-party conversational speech database collected at our research institute (ATR) [13]. The database contains multiple sessions of face-to-face conversations among three speakers (as the settings shown in Fig. 1). Audio, video, and motion data are available for each speaker. Each dialogue session includes 15 to 30 minutes of free-topic conversations. Each speaker wears a headset microphone (DPA4060), and a video-camera (of Kinect sensors) is set for each speaker for taking the frontal view.

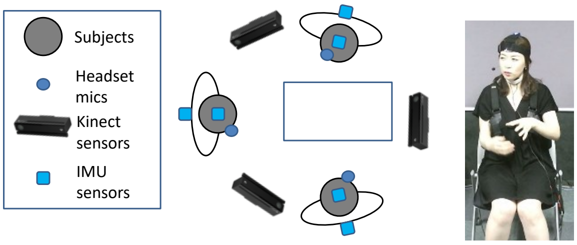

Figure 1: *Setup for three-party dialogue data collection, and an example of the image captured by one of the Kinect sensors.*

We used the data of 12 speakers (6 females and 6 males) appearing in 7 Japanese dialogue sessions (some of the speakers participated in multiple sessions). The female speakers are research assistants in their 30s and 40s; the male speakers are graduate students in their 20s.

2.1. Segmentation and annotation of eye movements

The eye gaze movements were segmented by a research assistant with experience in audio-visual data annotation, by carefully looking at gaze changes in the video data, frame-by-frame. The segmentation process resulted in a total of about 30,000 segments.

Then, the gaze segments were categorized in two main groups: one for gaze directed to a person's face, and another for gaze not directed to a person's face (gaze aversions). Fig. 2 shows the percentages in time of these two gaze patterns for

http://dx.doi.org/10.21437/Interspeech.2021-2134

female (F01 to F06) and male (M01 to M06) speakers. It can be observed that in this data set female speakers tend to look more at the dialogue participant's face, while male speakers tend to avert their gaze with higher frequency.

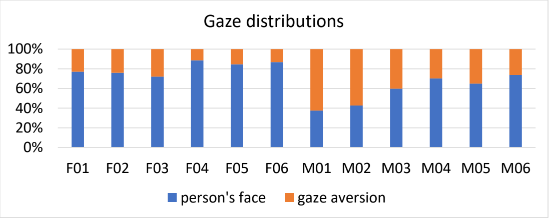

Figure 2: *Distributions (percentages is time) of gaze directed to a person's face and of gaze aversions, for each speaker.*

For the gaze segments not directed to a person's face, the position of the pupil was annotated to analyze gaze aversion behaviors in more detail. Annotation was conducted in terms of the following nine positions {left up, up, right up, left, front, right, left down, down, right down}, and an additional label in the case the eyes are closed. Although this process could be automated by using facial image processing tools, we preferred to conduct manually, since errors may occur when the face is turned to the left/right side directions, which often occur in the three-party scenario.

Further, to avoid human errors, the segmentation data was checked and fixed by another research assistant.

2.2. Annotation of gaze roles

The annotation of gaze roles was conducted for each eye motion segment according to different criteria, depending whether or not the gazing is directed to a dialogue partner.

Since no previous studies could be found providing a thorough guideline for annotation of gaze reasons, the annotators were asked to attribute gaze reason categories in a free-style. Consequently, the following category set has resulted for the gazing events directed to a dialogue partner. (The categories with low occurrence rates are omitted.)

- Talk: when the speaker's gaze is directed to the target **main listener** (main addressee). Two levels are annotated to distinguish when the speaker talks **enthusiastically**.
- Listen: when a listener's gaze is directed to the **speaker**. Two levels are annotated to distinguish when the listener expresses interest to the speaker's talk.
- Engage sub-listener: when the speaker sometimes gazes to a **sub-listener** (listeners other than the main listener), while talking to the main listener, to facilitate sub-listener's engagement.
- Check listener's reaction: when the speaker intends to check the listener's reaction on what the speaker said.
- Search next topic or next speaker: when the dialogue participants look at each other to decide who talks next, after a dialogue deadlock state.
- Predict next speaker: when the sub-listener gazes to the next speaker, before the speaker's utterance ends (usually at the end of a question directed to the next speaker, or at the end of an answer to a question made by the next speaker.)
- Wait response: when the gaze is directed to the next speaker, for waiting a response from the next speaker.
- Referred person: when gaze is directed to the person whom the speaker's gaze or pointing was directed to.
- React to other's behaviors: when gaze is reactively directed to the person's face, who made **backchannels** (verbal responses such as "uhm", "really"), **laughter**, **gestures** or **self-adapter** (self-touch) movements.
- Induce next speaker: when gaze is directed to a specific person to force him/her to take to floor.
- Topic sub-listener: when the dialogue participants gaze to a sub-listener, when the speaker's topic is about that sub-listener.

Next, for the gazing events not directed to a person's face. This includes gaze aversions to the environment and gaze directed to self or other's body parts (excluding the face).

- Gaze aversions when thinking
- Gaze aversions when remembering
- Gaze aversions during speaker's head motion: when speaker's gaze is averted (displaced from a dialogue partner's face) during his/her head motion (nodding, multiple nodding, head shaking, head tilting).
- Gaze aversions when laughing
- Gaze aversions for resting: when the gaze is thought to be averted to avoid excessive eye fixation toward a dialogue partner.
- Gaze aversions for quoted utterances: when the speaker utter a past or someone else's utterances.
- Gesture, adapters: when gaze is directed to self or other's gestures and adapters (self-touch) motions.
- Pointing target (excluding persons): when gaze is directed to a target object or direction pointed by a dialogue member.
- Gaze target (excluding persons): when gaze is directed to someone's gaze target, excluding the cases the gaze target is a dialogue member.
- External stimulus (sounds and light): when gaze is directed to an external stimulus, such as doors opening/closing, footsteps, intermittent noises and lights.

Fig. 3 and 4 shows the number of occurrences found for gaze segments directed to a person's face (upper panel) and for gaze aversion segments (lower panel), for different gaze reason categories. Gaze reason categories with low occurrences (less than 20) for both female and male speakers are omitted from the figures, and from subsequent analysis.

It can be observed from Fig. 3 that "listen" and "talk" (which are the basic expected behaviors during a dialogue interaction) shows the highest occurrences (with more than 1000 segments), followed by "predict next speaker", "wait next speaker", "check reaction" and "topic sub-listener" (with more than 100 segments), which are related to dialogue flow regulation. Occurrences close to 100 segments were also observed in "backchannel" and "laugh", which are related to reactions made by the listeners.

Regarding gaze aversions (Fig. 4), "thinking" and "remembering" have the highest occurrences (around or more than 1000 segments), followed by "nodding", "laughing", and "gesture(self)" (with more than 100 segments), which are related to own's non-verbal behaviors. For the dialogue setting in this study, referential gazing towards objects and reactive gazing to external stimulus were seldom observed.

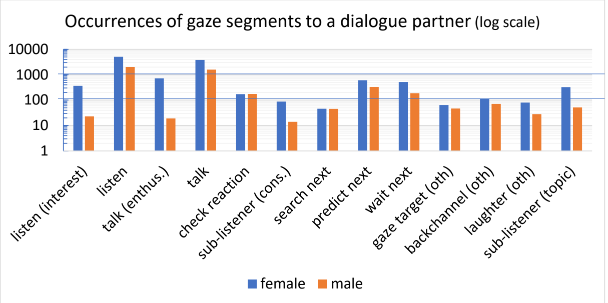

Figure 3: *Distributions of gaze segments directed to a person's face, for different gaze reason categories.*

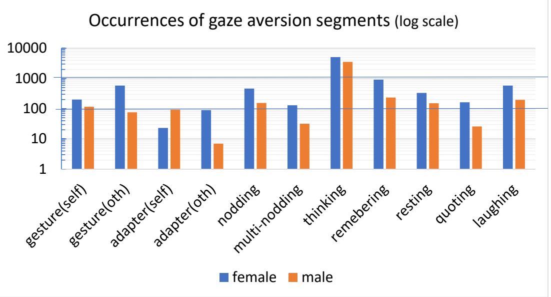

Figure 4: *Distributions of gaze aversion segments, for different gaze reason categories.*

3. Analysis results

In this section, we analyze gaze segment durations and gaze directions for each of the gaze reason categories introduced in Section 2.

3.1. Analysis of duration of gaze segments

Fig. 5 and 6 shows the average durations of the gaze segments directed to a dialogue partner, and gaze aversion segments, for different gaze reason categories. (The categories with low occurrences are omitted from the figures.)

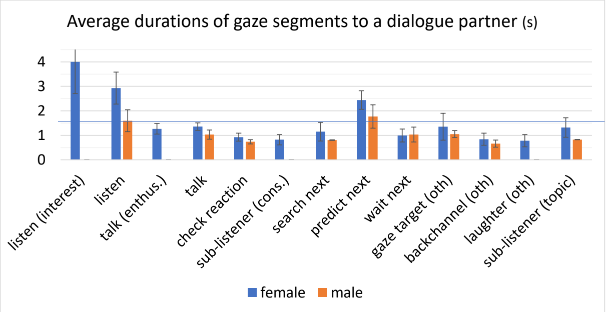

Figure 5: *Average durations of gaze segments directed to a person's face, for different gaze reason categories.*

Longer average durations above 1.5 seconds are found in listening mode ("listen"), and when the next speaker can be predicted in advance ("predict next"). In this dataset, female speakers tended to look longer toward the dialogue partners, in comparison to male speakers. For the "listen" mode in female speakers, the average looking duration is longer when they show interest to the speaker's talk ("listen(interest)"), which means gaze aversion becomes less frequent. In contrast, no differences are found in the speaking mode ("talk" and "talk(enthus.)"). This means that even when speakers talk enthusiastically, they do not fix their eyes at the listeners.

Short gazing with average durations around 0.5 to 1 second can be observed in "check reaction", "sub-listener", "search next", "backchannel(oth)" and "laughter(oth)", for both female and male speakers. This indicates that in such situations, a glancing look is made to acknowledge they perceived the other's reactions, or to increase engagement of the dialogue participants.

Regarding the results for gaze aversions shown in Fig. 6, it can firstly be observed that the average durations are mostly short, around 0.5 to 1 seconds.

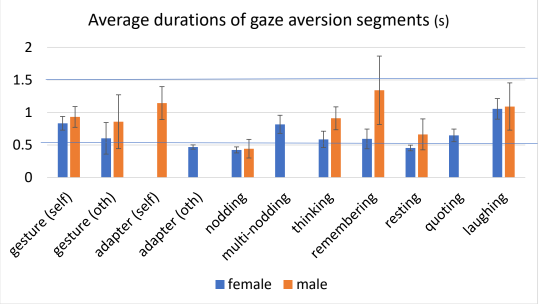

Figure 6: *Average durations of gaze aversion segments, for different gaze reason categories.*

In "thinking" and "remembering", shorter average durations are observed in female speakers, and longer average durations around 1 second or more are observed in male speakers. It is interesting to note that when speakers are thinking or remembering, they do not fix their gaze to a single point for a long time, but rather, the eyeballs frequently move around different directions (with average durations around 0.5 seconds per segment). Details on how the eyeballs move will be reported in the next subsection.

The average durations of "nodding", "multi-nodding", and "laughing" are dependent on the length of the accompanying single backchannels (responsive utterances such as "uhm"), multiple backchannels (such as "uhm uhm"), and laughter intervals, respectively.

Overall, female speakers tend to have longer average durations in gazing directed to a person, while male speakers show longer average durations in gaze aversions, in comparison to their counterparts.

3.2. Analysis of eyeball movements during gaze aversion

We then conducted analysis of the eyeball movements (i.e. the pupil position in the eye), during gaze aversion. Fig. 7 shows the distributions (occurrence rates) of the pupil positions, for different gaze aversion reason categories. (The gesture-related categories are excluded from this analysis, since the gazing directions are specified by the places the gestures occur. The distributions for multi-nodding are merged with the ones in nodding, since they showed similar trends.)

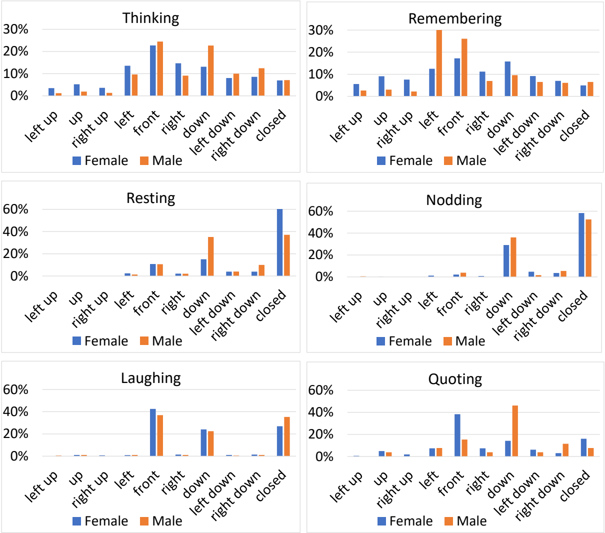

Figure 7: *Distributions of the pupil positions, for different gaze aversion reason categories.*

Differences and similarities can be observed in the gaze aversion patterns of different categories.

First, the distributions in “thinking” and “remembering” show gazing in several directions, with higher concentration in front and down directions. For remembering, higher occurrences are observed in left for male speakers, but this may be related to the positioning of the speaker relative to their main listener, who was predominantly in their right side.

“Resting” and “nodding” show higher occurrences in down and closed eyes. This can be explained by the fact nodding is a down-up head motion, so that there is tendency to be accompanied by a downward gazing. In “laughing”, gaze aversion distributions are concentrated in front, down, and closed eyes, which are likely to be coordinated with the head pose during laughter.

These results suggest that during nodding and laughing, the subject is already expressing a reaction to the speaker, so that they don’t need to keep eye contact all the time. Thus, this could be taken as a resting for eye contact.

Note that front direction is often observed in most of gaze aversion categories, since in this three-party dialogue scenario (Fig. 1), the dialogue partners are in the left and right sides, so that no one is in front of the speaker. It is worth mentioning that these distributions would be different for different positioning of the dialogue partners (specifically, if the participants are in front of each other).

Regarding the “quoting” category, it was interesting to find that the gaze aversions work as a visual cue for indicating that the utterance is a quotation, jointly with other auditory cues like changes in intonation and voice quality [14].

4. Conclusions and final remarks

We conducted analysis of the reasons/intentions behind the gaze behaviors of multiple speakers in three-party dialogue interactions. We categorized the reasons of gaze events directed to a person’s face, and of gaze aversions (i.e., gazing not directed to a person’s face). Detailed analysis on the eyeball movements during gaze aversions was also conducted.

Analysis of gazing directed to a dialogue participant’s face indicated that average gaze durations increase in listening mode when showing interest to the speaker’s talk, but do not change much in speaking mode even when talking enthusiastically. In contrast, short (glancing) looking (around 0.5 to 1 second) is predominant when a dialogue partner makes some reaction (such as backchannels, laughter, head and hand gestures), or when looking to a sub-listener in order to increase engagement of the dialogue participants.

Regarding gaze aversions, overall short gaze aversion segments (around 0.5 to 1 second) were found to be predominant, even during thinking and remembering. This means that even when the speakers are averting their gaze for a long period (during thinking and remembering), their eyes (pupils) are not constantly fixed to one direction, but rather are frequently moving with short intervals.

The gaze aversions during the speaker’s own non-verbal behaviors (such as when nodding and laughing) were found to have different patterns with those in thinking and remembering, being coordinated with the head motions.

Some gender differences were found in the analysis results of gazing in the present study. In general, female speakers tend to look more at their dialogue partner’s face, while male speakers tend to avert more their gaze. Considering that all male speakers available in our dataset were in their 20s, while all female speakers were in their 30s to 40s, it would be interesting to further investigate how factors like speaker’s age, personality, and cultural background influence in this gazing balance.

In future work, we also intend to apply the analysis results for automatically control the gaze behaviors of agents and robots in human-robot interactions.

5. Acknowledgements

This work was supported by JSPS, Grant Number JP20H05576, and by JST, Moonshot R&D Grant Number JPMJMS2011.

We thank Taeko Murase, Yuka Nakayama and Kyoko Nakanishi for their contributions on data annotation and analysis.

6. References

[1] A. Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, Vol. 26, pp. 22–63, 1967.

[2] M. Argyle and M. Cook. Gaze and mutual gaze. *Cambridge University Press*, 1976.

[3] K. Ruhland, C.E. Peters, S. Andrist, J.B. Badler, N.I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell. A review of eye gaze in virtual agents, social robotics and HCI: Behaviour generation, user interaction and perception. In Computer graphics forum, Vol. 34, pp. 299–326. *Wiley Online Library*, 2015.

[4] H. Admoni and B. Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, Vol. 6, No. 1, pp. 25–63, 2017.

[5] H. Sacks, E.A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pp. 7–55. Elsevier, 1978.

[6] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro. Conversational gaze mechanisms for humanlike robots. *ACM*

Transactions on Interactive Intelligent Systems (TiiS), Vol. 1, No. 2, pp. 1–33, 2012.

[7] G. Skantze, A. Hjalmarsson, and C. Oertel. Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication*, Vol. 65, pp. 50– 66, 2014.

[8] D. Lala, K. Inoue, P. Milhorat, and T. Kawahara. Detection of social signals for recognizing engagement in human-robot interaction. *arXiv preprin*t arXiv:1709.10257, 2017.

[9] M.A.A. Dewan, M. Murshed, and F. Lin. Engagement detection in online learning: a review. *Smart Learning Environments*, Vol. 6, No. 1, pp. 1–20, 2019.

[10] R. Ishii, K. Otsuka, S. Kumano, R. Higashinaka, and J. Tomita. Analyzing Gaze Behavior and Dialogue Act during Turn-Taking for Estimating Empathy Skill Level. Proceedings of *the 20th ACM Intl. Conf. on Multimodal Interaction*, p. 31-39, 2018.

[11] Y.I. Nakano, T. Yoshino, M. Yatsushiro, and Y. Takase. Generating robot gaze on the basis of participation roles and dominance estimation in multiparty interaction. *ACM Trans. Interact. Intell. Syst*., Vol. 5, No. 4, December 2015.

[12] S. Andrist, W. Collier, M. Gleicher, B. Mutlu, and D. Shaffer. Look together: Analyzing gaze coordination with epistemic network analysis. Frontiers in psychology, Vol. 6, p. 1016, 2015.

[13] C.T. Ishi, D. Machiyashiki, R. Mikata, H. Ishiguro. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters* 3(4), 3757–3764, July 2018.

[14] C.T. Ishi, H. Ishiguro, N. Hagita. Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech. *EURASIP Journal on Audio, Speech, and Music Processing 2010,* ID 528193, 1-12, Jan. 2010.