Comparing Video, Avatar, and Robot Mediated Communication: Pros and Cons of Embodiment

Kazuaki Tanaka[1,2], Hideyuki Nakanishi[1], and Hiroshi Ishiguro[3]

[1] Department of Adaptive Machine Systems, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
{tanaka,nakanishi}@ams.eng.osaka-u.ac.jp
[2] CREST, Japan Science and Technology Agency
[3] Department of Systems Innovation, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
ishiguro@sys.es.osaka-u.ac.jp

Abstract. In recent years, studies have begun on robot conferencing as a new telecommunication medium. In robot conferencing, people talk with a remote conversation partner through teleoperated robots which present the bodily motions of the partner with a physical embodiment. However, the effects of physical embodiment on distant communication had not yet been demonstrated. In this study, to find the effects, we conducted an experiment in which subjects talked with a partner through robots and various existing communication media (e.g. voice, avatar and video chats). As a result, we found that the physical embodiment enhanced social telepresence, i.e., the sense of resembling face-to-face interaction. Furthermore, the result implied that physical embodiment built the sense of tension as in the case of a first face-to-face meeting.

Keywords: Robot conferencing, physical embodiment, social telepresence, social anxiety, teleconferencing, videoconferencing, audio communication, avatar, face-to-face communication, face tracking.

1 Introduction

Currently, we can easily use audio and videoconferencing software. Audio-only conferencing, such as a voice chat, has a problem in that social telepresence decreases. The social telepresence is the sense of resembling face-to-face interaction [7]. Enhancing social telepresence psychologically makes the physical distance between remote people less and saves time and money on travel. The most common method of enhancing social telepresence is videoconferencing. It had been proposed that live video can transmit the social telepresence of a remote conversation partner [6, 11]. Nevertheless, videoconferencing does not transmit sufficient social telepresence compared with face-to-face conferencing.

To further enhance social telepresence, recent studies have begun on robot conferencing in which people talk with a remote conversation partner through teleoperated robots. The teleoperated robot uses motion tracking technologies to reflect partner's

T. Yuizono et al. (Eds.): CollabTech 2014, CCIS 460, pp. 96–110, 2014.

facial and body motions in real time. The main features of robot conferencing are to transmit conversation partner's body motions and to present these motions via a physical embodiment. The physical embodiment means the substitution of a partner's body that exists physically in the same place as a user. Thus, it is expected that the user can talk with feeling close to face-to-face. Some studies reported superiorities of robot conferencing to videoconferencing [15, 22]. One such study showed that the teleoperated robot which has a realistic human appearance enhances social telepresence compared with audio-only conferencing and videoconferencing [22]. Even so, it is difficult that each user owns a robot with his/her realistic appearance due to the high cost. For this reason, a teleoperated robot that has a humanlike face without a specific age or gender is developed [21]. However, it has not been yet found whether such an anonymous robot that transmits only body motions without disclosing an appearance still has superiority to videoconferencing.

As the communication medium similar to the robot conferencing, avatar chats are available. Recently, it has become easy and inexpensive to use avatar chats such as avatar Kinect. The avatar chat resembles the robot conferencing in transmitting user's body motions without disclosing the user's appearance, but differs in reflecting these movements onto a computer graphics animation which does not have a physical embodiment. A lot of studies found positive effects of avatar on distant communication [2, 4, 8, 12, 25]. Several such studies reported that avatar chats enhance social telepresence compared with audio-only conferencing [4, 12]. In addition, it was also reported that avatars increase the degree of smoothness of speech as well as video conferencing [25]. Thus, the transmitting body motions might be enough to produce these positive effects. If the physical embodiment does not produce such positive effects, the usefulness of robot conferencing would decrease since robots are more expensive than videos and avatars.

Therefore, to prove that robot conferencing is useful, it is necessary to demonstrate that the physical embodiment improves distant communication independently from the transmitting information, e.g., audio, motion and appearance. In this study, we investigated how the physical embodiment and transmitting information factors influence the social telepresence and the degree of smoothness of speech. To analyze the effects of the two factors separately, we prepared six communication methods as shown in Fig. 1. The voice chat, avatar chat and videoconferencing that do not have a physical embodiment transmit audio only, audio + motion and audio + motion + appearance respectively. The robot conferencing that has a physical embodiment transmits audio + motion, and so it corresponds to the avatar chat as described above. As the method that corresponds to the voice chat, we set an inactive robot conferencing that transmits audio but no motion. Furthermore, we assumed that the face-to-face communication corresponds to the videoconferencing.

2 Related Work

There are many studies related to robot conferencing. They have proposed various teleoperated robots that present the operator's facial movements [13, 15, 21, 22, 24] with a physical embodiment. A previous study that evaluated robot conferencing with

regard to social telepresence concluded that robot conferencing transmitted a higher social telepresence of the remote conversation partner than audio-only and videoconferencing [22]. However, the teleoperated robot that was used in the study had a specific person's appearance, and so it was not clear which of the factors, the physical embodiment or the appearance, enhanced the social telepresence. Additionally, the teleoperated robot reproduced the whole body of a person, whereas the videoconferencing only showed conversational partner's head. The video image of only a head is harmful to social telepresence [20], so that a superiority of robot conferencing to videoconferencing which shows the whole body of a person was also not clear. To clarify them, we used an anonymous teleoperated robot [21] that has a humanlike face without a specific age or gender, and compared it with life-size communication media that reproduced the whole body of the conversational partner.

In videoconferencing research, it was reported that the remote person's movement that was augmented by a display's physical movement enhanced the social telepresence [18]. Furthermore, in the human-robot interaction field, there are studies that focused on the effects of the physical embodiment of robot agents on social presence [3, 14]. These studies showed that people feel higher social presence from robot agents than on-screen agents. There is a possibility that people also feel higher social telepresence during robot conferencing due to the effects of the physical embodiment.

The superiority of robot conferencing to videoconferencing was indicated also in objective measures. One such study showed that the eye-gaze of remote person reproduced by a robot was more recognizable than by a live-video [15]. Most previous studies that tried to find influences of a difference between communication media on objective measures have focused on conversational structures such as turn-taking and overlapping [1, 5, 23]. In this study, we observed the frequency of pauses and percentage of pause times in speech to measure the degree of smoothness of speech [25]. The concrete methods to calculate them are explained in Section 3.4.

The previous studies that showed superiorities of robot conferencing dealt with each telecommunication media as a single factor [15, 22]. By contrast, this study divided the telecommunication media into a physical embodiment factor and a transmitting information factor. For example, we assumed that a robot has a physical embodiment and transmits audio and body motions, and a video does not have a physical embodiment and transmits audio, body motions and appearances.

3 Experiment

3.1 Hypothesis

In this study, we conducted an experiment to confirm how features of robot conferencing influence distant communication. The main features of robot conferencing are to have a physical embodiment and to transmit conversation partner's body motions. We predicted that these features enhance social telepresence, and so we made the following two hypotheses.

Hypothesis 1: A physical embodiment enhances the social telepresence of the conversation partner.
Hypothesis 2: Transmitting body motions enhances the social telepresence of the conversation partner.

The previous study that investigated the influence of difference between communication media on the degree of smoothness of speech showed that body motions presented by videos and avatars decreased speech pauses compared with audio-only media [25]. In addition, it was reported that videos and avatars enhance social telepresence compared with audio-only media [4, 12]. On the assumption that enhancing social telepresence smoothens speeches, we predicted that the features of robot conferencing could decrease speech pauses. Thus, we added the following two hypotheses.

Hypothesis 3: A physical embodiment smoothens a speech that is directed to the remote conversation partner.
Hypothesis 4: Transmitting body motions smoothens a speech that is directed to the remote conversation partner.

3.2 Conditions

The hypotheses described in the preceding section consist of these two factors: physical embodiment and transmitting information. The physical embodiment factor had two levels, with/without physical embodiment, and the transmitting information factor had three levels, audio, audio + motion and audio + motion + appearance. Thus, to examine the hypotheses, we prepared six conditions of a 2x3 design shown in Fig. 1.

As described in Section 1, both robot conferencing and avatar chat transmit remote person's body motions without disclosing the person's appearance. We thus supposed that the avatar chat can become robot conferencing by adding a physical embodiment. Similarly, we assumed that the voice chat becomes an inactive robot conferencing which does not transmit the body motions of a remote person and the video chat can become face-to-face communication by adding a physical embodiment. In terms of the transmitting information, we assumed that the voice chat and inactive robot transmit only audio, the avatar and robot transmit audio and motion, and the video and face-to-face transmit audio, motion and appearance. These assumptions allowed us to analyze the effect of adding a physical embodiment to existing communication media. The details of each condition are described below.

Active Robot Condition (Transmitting Audio and Motion with a Physical Embodiment): The subject talked to the conversation partner while looking at the robot. The robot had a three-degrees-of-freedom neck and a one-degree-of-freedom mouth. The head and lips moved at thirty frames per second according to the sensor data sent from face tracking software (faceAPI), that was running in a remote terminal and capturing the conversation partner's movements. The camera for face tracking was set behind the robot. The microphone speaker was set behind the robot. The robot was dressed with the same gray shirt as the conversation partner.

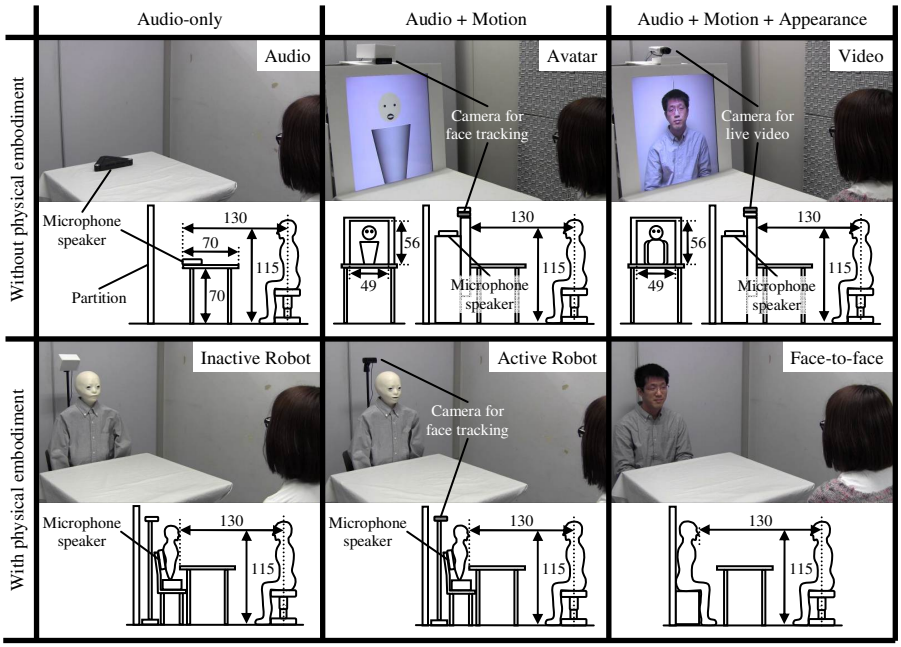

Fig. 1. Experimental conditions and setups (length unit: centimeters)

Avatar Condition (Transmitting Audio and Motion But No Physical Embodiment): The subject talked to the conversation partner while looking at an anonymous three-dimensional computer graphics avatar that reflected the conversation partner's head and lip motions. The avatar consisted of a skin-colored cylindrical head, black lips, black eyeballs and a gray conical body which was the same color as the shirt of the conversation partner. In the preliminary experiment, we used an avatar which had a spherical head and a realistic shirt, which looked like the robot. However there were some subjects who felt hard to notice facial movements of the avatar. This problem was solved by changing the design of avatar to a cylindrical head. The recognizable facial movements might improve social telepresence, and so we employed the cylindrical head. In addition, we modified its body to a conical shape to standardize the abstraction level of the looks. The diameter of the head was equal to the breadth of the robot's head (13.5 cm). The conversation partner's head and lip motions were tracked in the same way as on the active robot condition. The head translated and rotated with three degrees of freedom. The lips were transformed based on the three-dimensional positions of fourteen markers. The head and lips moved at thirty frames per second. The avatar was shown on a 40-inch display. The display was set longitudinally on the other side of the desk. The bezel of the display was covered with a white board, so that the true display area was 49 cm by 56 cm. The microphone speaker was set behind the display. There were two cameras on top of the display. One was for face tracking, and the other was for live video. In this condition, the camera for live video

was covered with a white box. The camera was used in the video condition described below.

Face-to-Face Condition (Transmitting Audio, Motion and Appearance with a Physical Embodiment): The subject talked to the conversation partner in a normal face-to-face environment. The conversation partner wore a gray shirt. The distance from the subject to the conversation partner was adjusted to 150 cm so that the breadth of the conversation partner's head looked the same as the breadth of the robot's head (13.5 cm).

Video Condition (Transmitting Audio, Motion and Appearance But No Physical Embodiment): This condition was identical to a normal video chat. The subject talked to the conversation partner while looking at a live video of the conversation partner. The conversation partner wore a gray shirt. The resolution of the camera for live video was 1280 pixels by 720 pixels, and its frame rate was 30 frames per second. The video was shown on the same display that was used in the avatar condition. Thus, the true display area was 49 cm by 56 cm. The horizontal angle of view was adjusted to 87 degrees so that the breadth of the conversation partner's head was equal to the breadth of the robot's head (13.5 cm) on the display. The camera for face tracking that was used on the avatar condition was covered with a white box.

Inactive Robot Condition (Transmitting Audio with a Physical Embodiment): The subject talked to the conversation partner while looking at the inactive robot. The camera for face tracking that was used on the active robot condition was covered with a white box. The subject was preliminarily informed that the robot did not move in this condition.

Audio-Only Condition (Transmitting Audio But No Physical Embodiment): This condition was similar to a normal voice chat. The subject talked to the conversation partner through only a microphone speaker that was set on the desk.

In the preliminary experiment, some subjects doubted that the experimenter would be looking at them from somewhere even if the experimental condition required no camera. We hence informed the subjects that the dialogue environments of the subject side and the conversation partner side were the same in all the conditions. To make the subjects believe this bi-directionality of the dialogue environments, the subjects were shown a live video of the subjects' avatar, robot or video which were seen by the conversation partner on a 7-inch display before each experiment. At the same time, the subjects confirmed that their avatar and robot reflected their face and lip movements. The subjects also confirmed that the avatar and robot in front of them reflected the conversation partner's face and lip movements by comparing a live video of the conversation partner that was shown on the 7-inch display with the avatar and robot. The 7-inch display for these confirmations was removed before the experiments.

3.3 Task

In the experiment, the subject talked with the conversation partner in the six conditions described above. An experimenter played the role of the partner. To observe the

difference in the social telepresence between the conditions, we asked the subject to answer a questionnaire (which is explained in the next section) after the experiment ended. Additionally, we observed the speech pauses. It therefore was necessary to record extended speech, which was required to stably measure the speech pauses. Simultaneously, it was also necessary to avoid analyzing speech that was interleaved with a remote conversation partner's replies to the subject, because those replies would become noise that affected the following utterances of the subject. To collect such speech, we created a task in which the subject could continue to talk for more than one minute without the partner's interference. While the subject was talking, the experimenter did not talk and gave only minimum backchannel responses with an utterance and a small nod of his head.

The subject was asked by the experimenter to talk about the issue and resolution of a certain gadget and requests for a new function on that gadget at the beginning of each condition. Because all the subjects had to experience the six conditions, we prepared six gadgets as conversational topics, i.e., e-book readers, handheld game consoles, smartphones, robotic vacuum cleaners, portable audio players, and 3D televisions. We did not disclose the next topic beforehand, and the experimenter told the subject which gadget to talk about right when the condition began.

We did not ask the subject to talk for more than a certain duration, so the subject could stop talking anytime. However, since the six gadgets are attracting considerable attention recently, most subjects knew the issue and resolution of the gadgets to a certain level, and their speech was able to last more than one minute. A one-minute speech would be too short to analyze turn-taking, but it was enough to observe the difference in the pauses.

The order of experiencing the conditions and the order of the topics were counterbalanced. The subject trained the task in the face-to-face condition in order to familiarize the subject with the task and the experimenter's motion and appearance, before conducting the experiment in the six conditions. The topic of the training was always railway smart cards.

3.4 Data Collection

Questionnaire

After experiencing the six conditions, the subjects answered a questionnaire, which asked them to estimate the social telepresence, i.e., the degree of resembling face-to-face interaction [7] for each condition. The questionnaire is shown in Fig. 2. The questionnaire had six statements that corresponded to the six conditions. The statement was the following: I felt as if I were talking to the conversation partner in the same room. Previous studies showed that the statement which asks a feeling of being in the same room is useful to measure the social telepresence [16, 17, 18, 19]. The statement was rated on a 9-point Likert scale where 1 = strongly disagree, 3 = disagree, 5 = neutral, 7 = agree, and 9 = strongly agree. The subjects thereby could score the same number on the statements if they felt the same level of social telepresence in the conditions.

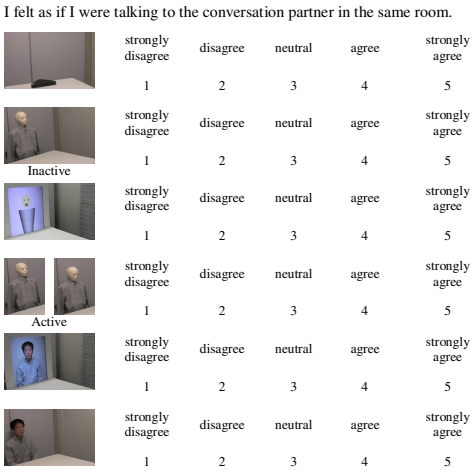

Fig. 2. Questionnaire to evaluate the social telepresence

The statements were sorted in the order of the conditions and were printed on the questionnaire, with a photo that showed the experimental setup of the corresponding condition. The sort and the photo were good cues to help the subjects remember the feeling of social telepresence in each condition. After answering the questionnaire, the subjects were interviewed. The interview was conducted in order to ask the subjects the reason of scoring of the questionnaire.

Speech Pauses

To calculate the frequency of pauses and the percentage of pause times in the recorded speech, the speech was transcribed by using a multimedia annotation tool, ELAN, which can partially repeat recorded speech and show its waveform. The waveform was good cue to discriminate between pause parts and utterance parts. Three experimenters used this tool for the transcription and the pause estimation. We entered the beginning time, ending time, and transcript of all the utterances to count the pauses.

To exclude arbitrariness from the analysis, we did not have any minimum or maximum threshold for the length of a pause to be counted. However, we could not count a pause that was shorter than fifty milliseconds, because it was actually impossible to distinguish such a short pause from a speaking part due to white noise. We did not filter out white noise, because the filtering could also cut off utterances spoken very quietly. The quiet utterances sometimes mixed with white noise. In such a case, we listened to the part over and over again, shifting the beginning and ending time of the part in steps of ten milliseconds. As a result, we estimated pause parts and utterance parts as accurately as possible. The transcription which was made by one person was checked by two people to confirm the consistency of the pause estimation. In addition, when there were discrepancies in the pause estimation among us, we had ample discussions about it.

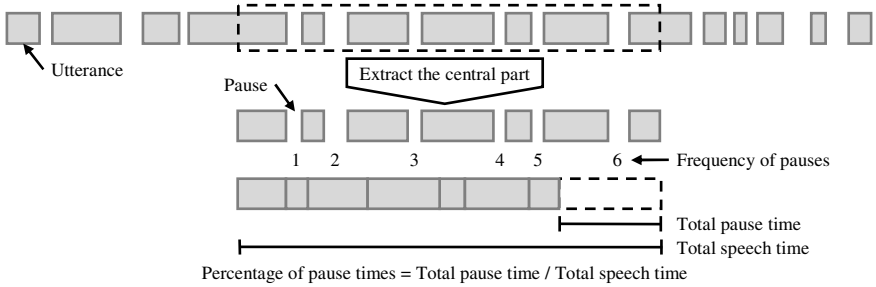

Fig. 3. Method to calculate the frequency of pauses and the percentage of pause times

Fig. 3 shows the method for calculating the frequency of pauses and the percentage of pause times. First, we extracted the central part of the speech. The speech of most subjects lasted more than one minute, which corresponded to about two-hundred syllables in our language (Japanese). Thus, we extracted the central two hundred syllables of the speech for the analysis. This extraction equalized the amount of speech data across conditions and subjects. This extraction also stabilized the analysis, since the pauses of the beginning or ending part of the speech was affected by individual subjects. The beginning part tended to be unfairly smooth if the subject was accidentally ready to talk about the gadget. For example, one subject was considering the purchase of the gadget. Further, the ending part tended to needlessly increase the pauses if the subject made an extra effort to continue talking.

Next, we counted the number of pauses included in the central part. The frequency of pauses is the number. And, we calculated the total speech time and total pause time in the central part. The percentage of pause times was calculated by dividing the total pause time by the total speech time. We used a spreadsheet software to count the number of syllables, extract the central part, count the number of pauses, and calculate the total speech time and total pause time.

4 Result

Thirty-six undergraduate students who lived near our university campus participated in the experiment and talked about the gadgets in the six conditions. Thus, we collected thirty-six questionnaires and 216 recorded speech in total. We analyzed their answers of the questions and speech pauses. The results of analysis are described below.

4.1 Social Telepresence

Fig. 4 shows the result of the questionnaire, in which each point represents the mean value of the scores, and each bar represents the standard error of the mean value.

We compared the six conditions to find the effects of the physical embodiment and the transmitting information factors. Since the physical embodiment and the transmitting information factors consisted of two and three levels as shown in Fig. 1 and each subject evaluated all conditions, we conducted 2x3 two-way repeated-measures

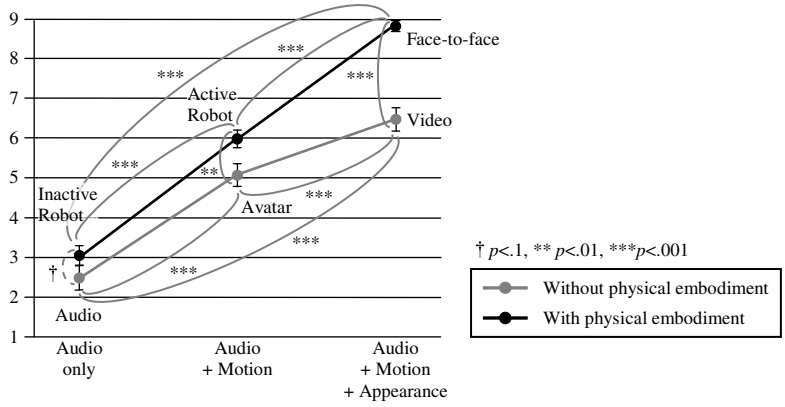

Fig. 4. Results of the questionnaire on the feeling of speaking to the partner in the same room

ANOVA. As a result, we found strong main effects of the physical embodiment factor ($F(1, 35)=36.955$, $p<.001$) and the transmitting information factor ($F(2, 70)=279.603$, $p<.001$). We also found a strong interaction between these factors ($F(2, 70)=14.794$, $p<.001$). We further analyzed the simple main effects in the interaction with the Bonferroni correction. The physical embodiment significantly improved the social telepresence of the conversation partner, when the transmitting information was audio + motion + appearance ($F(1, 105)=8.857$, $p<.01$), and audio + motion ($F(1, 105)=65.470$, $p<.001$). When the transmitting information was audio only, there was a non-significant tendency for the social telepresence to increase ($F(1, 105)=3.460$, $p=.086$). This meant that the subjects felt a higher social telepresence of the conversation partner in the face-to-face condition than in the video condition, and the active robot condition conveyed a higher social telepresence than the avatar. These results support hypothesis 1 that the physical embodiment enhances the social telepresence of the conversation partner. However, the effect of the physical embodiment on the social telepresence was low in the audio only communication.

Furthermore, there were significant differences between the three levels of the transmitting information in both cases of without physical embodiment ($F(2, 140)=223.095$, $p<.001$) and with physical embodiment ($F(2, 140)=107.141$, $p<.001$). Multiple comparisons showed that the subjects felt a higher social telepresence in the face-to-face condition than in the active robot ($p<.001$) and inactive robot ($p<.001$) conditions, the active robot condition conveyed a higher social telepresence than the inactive robot condition ($p<.001$), the video condition conveyed a higher social telepresence than the avatar ($p<.001$) and the audio-only ($p<.001$) conditions, and the avatar condition conveyed a higher social telepresence than the audio-only condition ($p<.001$). These results prove hypothesis 2 that transmitting body motions enhances the social telepresence of the conversation partner. In addition, transmitting appearance also enhances the social telepresence of the conversation partner.

4.2 Smoothness of Speech

In the experiment, most subjects could continue talking for more than one minute, but 10 subjects could not in a few conditions. To calculate the frequency of pauses and the percentage of pause times, we had to extract the central part of the speech because the beginning or ending part of the speech was affected by individual subjects as described in Section 3.4. The speech of less than one minute was too short to extract the

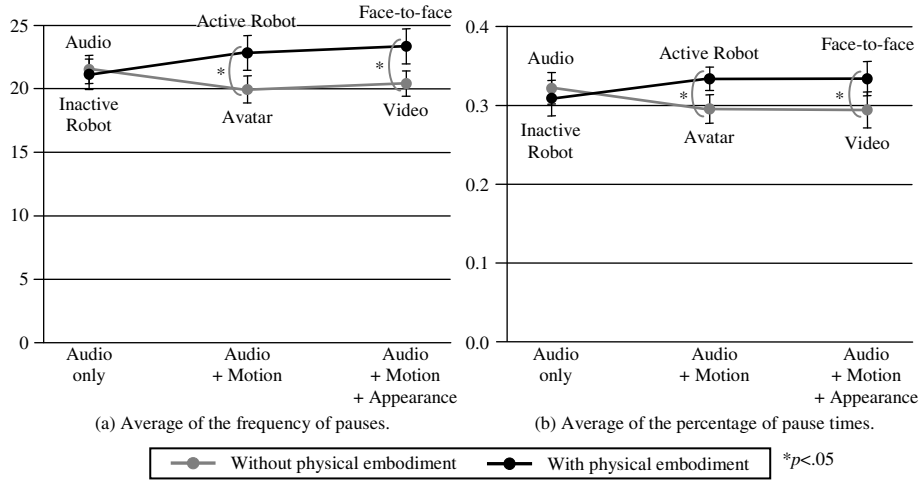

Fig. 5. Results of analyzing the speech pauses

central part. We therefore analyzed the pauses of twenty-six subjects who spoke for more than one minute in all the conditions.

Fig. 5(a) and (b) shows the mean value of the frequency of pauses and the percentage of pause times respectively, in which each point represents the mean value. To find the effects of the physical embodiment and the transmitting information factors on the frequency of pauses and the percentage of pause times, we conducted 2x3 two-way repeated-measures ANOVA in the same way as in the result of the questionnaire.

As a result of analyzing the frequency of pauses, we found a strong main effect of the physical embodiment factor ($F(1, 25)=8.004$, $p<.01$), but the main effect of the transmitting information factor was not significant. We also found a weak interaction between these factors ($F(2, 50)=2.947$, $p=.062$). We then analyzed the simple main effects in the interaction with the Bonferroni correction. The physical embodiment significantly increased the frequency of pauses, when the transmitting information was audio + motion + appearance ($F(1, 75)=6.981$, $p<.05$), and audio + motion ($F(1, 75)=6.799$, $p<.05$). There was no other significant effect of the physical embodiment factor.

As a result of analyzing the percentage of pause times, we found a weak main effect of the physical embodiment factor ($F(1, 25)=3.174$, $p=.087$), but the main effect of the transmitting information factor was not significant. We also found a weak interaction between these factors ($F(2, 50)=3.146$, $p=.052$). We then analyzed the simple

main effects in the interaction with the Bonferroni correction. The physical embodiment significantly increased the percentage of pause times, when the transmitting information was audio + motion + appearance ($F(1, 75)=4.647$, $p<.05$), and audio + motion ($F(1, 75)=4.369$, $p<.05$). There was no other significant effect of the physical embodiment factor.

These results meant that the subjects' speech included more pauses in the face-to-face condition than in the video condition, and the speech in the active robot condition had more pauses than in the avatar condition, against the hypothesis 3 that a physical embodiment smoothens a speech that is directed to the remote conversation partner. However, the physical embodiment did not influence speech pauses in audio-only communication. Additionally, the transmitting information also did not influence speech pauses, and so the hypothesis 4 that Transmitting body motions smoothens a speech that is directed to the remote conversation partner.

5 Discussion

In the experiment, the physical embodiment enhanced the social telepresence of the conversation partner. In the interviews, seven of the thirty-six subjects said that they felt as if they were facing the conversation partner in the active robot condition compared with the avatar condition because there was a physical object in front of them. However, there was no significant difference between the audio-only condition and the inactive robot condition. In the interviews, three of the thirty-six subjects said that the inactive robot condition was not that different to the audio-only condition because they could not see the conversation partner's reaction. In fact, eight of the thirty-six subjects scored the same number for the audio-only and inactive robot conditions in the questionnaire. Moreover, five of the thirty-six subjects said that they felt as if the conversation partner was in front of them when the robot moved. These subjective responses support the experimental result that a physical embodiment enhances social telepresence when transmitting body motions. This result indicates the superiority of robot conferencing to avatar chat which does not have a physical embodiment.

Presence or absence of motion parallax can be cited as one of the differences between physical embodiment and video. In robot conferencing, the depth from motion parallax could increase visibility of body motions. The lack of the depth information might be the cause of feeling hard to notice facial movements of the avatar used in the preliminary experiment described in Section 3.2. A previous study reported that motion parallax generated by the movement of a camera enhances social telepresence [17]. The visibility of bodily motion improved by the motion parallax may have contributed to enhance social telepresence.

In terms of the transmitting information, the appearance enhanced social telepresence of the conversation partner as well as the body motions. This result shows the disadvantage of robot conferencing and avatar chat that do not transmit the partner's appearance. Although the active robot has this disadvantage, the active robot and video conditions seemed to convey the same degree of social telepresence, as shown in Fig. 4. In the questionnaire, more than half of subjects (sixteen of the thirty-six) scored the same or higher number for the active robot condition than the video condition. We assumed that the enhanced social telepresence by the physical embodiment

offset the decreased social telepresence by the absence of the partner's appearance. Therefore, the reported superiority of robot conferencing in the social telepresence to video conferencing [22] could be caused by the robot's realistic appearance.

We predicted that enhancing social telepresence could increase the degree of smoothness of speech. However, against this prediction, the physical embodiment that enhanced social telepresence decreased the degree of smoothness. On the other hand, in Fig. 5(a) and (b), you can see that the frequency of pauses and the percentage of pause times of the avatar and video conditions seemed to be lower than that of the audio-only condition, although the difference was not significant. This could be caused by the effect of decreasing speech pauses by transmitting body motions that was reported in the previous study [25]. We are currently investigating the cause of the non-significance and anticipate that the several differences of experimental environments influenced the speech pauses, e.g. this study used the life-size avatar and video, whereas the previous study used the small-size of them.

The increasing of speech pauses by the physical embodiment might be caused by the sense of tension as in the case of a first face-to-face meeting. In the interviews, sixteen of the thirty-six subjects referred to a sense of tension. Fifteen of the sixteen subjects felt tension in the face-to-face condition, and five of the sixteen subjects felt tension in the active robot condition. By contrast, the subjects who felt tension in the video and avatar conditions were only two of the sixteen subjects respectively, and no subjects felt tension in the inactive robot and audio-only conditions. These subjective responses showed that the subjects felt tension when they could see the conversation partner's motions with a physical embodiment. In the social psychology field, it is known that social anxiety that is the uncomfortable feeling while talking with a conversation partner increases the frequency of pauses [10] and percentage of pause times [9]. We hence considered that the speech pauses in the face-to-face and active robot conditions were increased by social anxiety. The sense of tension when talking with a stranger is one of the social anxieties. Therefore, there is a possibility that robot conferencing builds a sense of tension as in the case of a first face-to-face meeting.

In this study, we did not investigate the conditions that transmit audio and appearance but not motion. Talking through an inactive robot that has a realistic appearance of a partner, and a partner's photo could correspond to such conditions. Watching the partner's photo while talking is a popular situation since many users of instant messengers put their photos in the buddy list. Although the transmitting appearance enhances social telepresence as mentioned above, it has not been clarified whether the appearance works even if the motion is not transmitted. By contrast, the effect of appearance on the smoothness of speech had already demonstrated [25]. The previous study showed that presenting the partner's photo did not increase the degree of smoothness of speech. We hence predict that the appearance also does not enhance social telepresence if the motion is not transmitted as is the case with the physical embodiment. To prove this hypothesis is a future work.

6 Conclusion

In this study, to investigate how the features of robot conferencing influence remote communication, we compared robot conferencing with existing communication media

divided into physical embodiment and transmitting information factors. We found that transmitting body motions via the physical embodiment enhances social telepresence. This result shows the superiority of robot conferencing to avatar chat. However, it was also found that presenting conversational partner's appearance which is not transmitted by robot conferencing enhances social telepresence. Consequently, robot conferencing was comparable to videoconferencing since the positive effect of the physical embodiment offset the negative effect of lacking appearance.

Previous studies have discussed the superiority of robot conferencing to videoconferencing. However, we conclude that robot conferencing in the absence of presenting remote person's appearance does not always have the superiority in social telepresence.

In addition, we analyzed the subjects' speech to examine how the physical embodiment and transmitting information factors affect the degree of smoothness of speech. As a result, we also found that transmitting body motions via the physical embodiment increases pauses in speech. This result implies the possibility that robot conferencing builds a sense of tension as in the case of the first face-to-face meeting because the increasing of pauses in speech might be caused by the sense of tension. Thus, robot conferencing could be suitable for interactions that require a sense of tension, e.g., interviews and lectures.

Acknowledgments. This study was supported by JSPS Grants-in-Aid for Scientific Research No. 21680013"Telerobotic media for supporting social telepresence", No. 20220002 "Representation of human presence by using tele-operated androids", JST CREST "Studies on Cellphone-type Teleoperated Androids Transmitting Human Presence" and Global COE Program "Center of Human-friendly Robotics Based on Cognitive Neuroscience."

References

1. Anderson, A.H., Newlands, A., Mullin, J., Fleming, A., Doherty-Sneddon, G., Van Der Velden, J.M.: Impact of Video-Mediated Communication on Simulated Service Encounters. Interacting with Computers 8(2), 193–206 (1996)
2. Bailenson, J.N., Yee, N., Merget, D., Schroeder, R.: The Effect of Behavioral Realism and Form Realism of Real-Time Avatar Faces on Verbal Disclosure, Nonverbal Disclosure, Emotion Recognition, and Copresence in Dyadic Interaction. Presence: Teleoperators & Virtual Environments 15(4), 359–372 (2006)
3. Bainbridge, W.A., Hart, J., Kim, E.S., Scassellati, B.: The benefits of interactions with physically present robots over video-displayed agents. International Journal of Social Robotics 1(3), 41–52 (2011)
4. Bente, G., Ruggenberg, S., Kramer, N.C., Eschenburg, F.: Avatar-Mediated Networking: Increasing Social Presence and Interpersonal Trust in Net-Based Collaborations. Human Communication Research 34(2), 287–318 (2008)
5. Daly-Jones, O., Monk, A.F., Watts, L.: Some Advantages of Video Conferencing over High-quality Audio Conferencing: Fluency and Awareness of Attentional Focus. International Journal of Human-computer Studies 49(1), 21–58 (1998)

6. de Greef, P., Ijsselsteijn, W.: Social Presence in a Home Tele-Application. CyberPsychology & Behavior 4(2), 307–315 (2001)
7. Finn, K.E., Sellen, A.J., Wilbur, S.B.: Video-Mediated Communication. Lawrence Erlbaum Associates (1997)
8. Garau, M., Slater, M., Bee, S., Sasse, M.A.: The Impact of Eye Gaze on Communication Using Humanoid Avatars. In: Proc. CHI 2001, pp. 309–316 (2001)
9. Goberman, A.M., Hughes, S., Haydock, T.: Acoustic characteristics of public speaking: Anxiety and practice effects. Journal of Speech Communication 53(6), 867–876 (2011)
10. Harrigan, J.A., Suarez, I., Hartman, J.S.: Effect of Speech Errors on Observers' Judgments of Anxious and Defensive Individuals. Journal of Research in Personality 28(4), 505–529 (1994)
11. Isaacs, E.A., Tang, J.C.: What Video Can and Can't Do for Collaboration: A Case Study. Multimedia Systems 2(2), 63–73 (1994)
12. Kang, S., Watt, J.H., Ala, S.K.: Communicators' Perceptions of Social Presence as a Function of Avatar Realism in Small Display Mobile Communication Devices. In: Proc. HICSS 2008(2008)
13. Kuzuoka, H., Yamazaki, K., Yamazaki, A., Kosaka, J., Suga, Y., Heath, C.: Dual Ecologies of Robot as Communication Media: Thoughts on Coordinating Orientations and Projectability. In: Proc. CHI 2004, pp. 183–190 (2004)
14. Lee, K.M., Jung, Y., Kim, J., Kim, S.R.: Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human-robot interaction. International Journal of Human-Computer Studies 64(10), 962–973 (2006)
15. Morita, T., Mase, K., Hirano, Y., Kajita, S.: Reciprocal Attentive Communication in Remote Meeting with a Humanoid Robot. In: Proc. ICMI 2007, pp. 228–235 (2007)
16. Nakanishi, H., Murakami, Y., Nogami, D., Ishiguro, H.: Minimum Movement Matters: Impact of Robot-Mounted Cameras on Social Telepresence. In: Proc. CSCW 2008, pp. 303–312 (2008)
17. Nakanishi, H., Murakami, Y., Kato, K.: Movable Cameras Enhance Social Telepresence in Media Spaces. In: Proc. CHI 2009, pp. 433–442 (2009)
18. Nakanishi, H., Kato, K., Ishiguro, H.: Zoom Cameras and Movable Displays Enhance Social Telepresence. In: Proc. CHI 2011, pp. 63–72 (2011)
19. Nakanishi, H., Tanaka, K., Wada, Y.: Remote Handshaking: Touch Enhances Video-Mediated Social Telepresence. In: Proc. CHI 2014, pp. 2143–2152 (2014)
20. Nguyen, D.T., Canny, J.: More than Face-to-Face: Empathy Effects of Video Framing. In: Proc. CHI 2009, pp. 423–432 (2009)
21. Ogawa, K., Nishio, S., Koda, K., Balistreri, G., Watanabe, T., Ishiguro, H.: Exploring the Natural Reaction of Young and Aged Person with Telenoid in a Real World. Journal of Advanced Computational Intelligence and Intelligent Informatics 15(5), 592–597 (2011)
22. Sakamoto, D., Kanda, T., Ono, T., Ishiguro, H., Hagita, N.: Android as a Telecommunication Medium with a Human-like Presence. In: Proc. HRI 2007, pp. 193–200 (2007)
23. Sellen, A.J.: Remote Conversations: The Effects of Mediating Talk with Technology. Human-Computer Interaction 10(4), 401–444 (1995)
24. Sirkin, D., Ju, W.: Consistency in physical and on-screen action improves perceptions of telepresence robots. In: Proc. HRI 2012, pp. 57–64 (2012)
25. Tanaka, K., Onoue, S., Nakanishi, H., Ishiguro, H.: Motion is Enough: How Real-Time Avatars Improve Distant Communication. In: Proc. CTS 2013, pp. 465–472 (2013)