Motion is Enough: How Real-Time Avatars Improve Distant Communication

Kazuaki Tanaka
Dept. of Adaptive Machine Systems
Osaka University
CREST, JST
Suita, Osaka, Japan
tanaka@ams.eng.osaka-u.ac.jp

Satoshi Onoue, Hideyuki Nakanishi
Dept. of Adaptive Machine Systems
Osaka University
Suita, Osaka, Japan
{satoshi.onoue, nakanishi}@ams.eng.osaka-u.ac.jp

Hiroshi Ishiguro
Dept. of Systems Innovation
Osaka University
Toyonaka, Osaka, Japan
ishiguro@sys.es.osaka-u.ac.jp

***Abstract*—Thanks to the maturity of motion tracking technologies, it became easy and inexpensive to use avatars that reflect the user's facial and body movements in real time. There is a possibility that those real-time avatars can serve as a substitute for videos in distant communication. We conducted an experiment to observe differences between video chat and voice chat, and examined whether avatar chat can produce the same differences. In the experiment the subjects watched a conversation partner's video, avatar, photo, or nothing when speaking to the partner. Since the video and avatar delivered the partner's motion while the video and photo delivered the partner's appearance, we were able to observe the effects of the motion and appearance separately. As a result, we found that presenting the motion via a video or an avatar increased the degree of the smoothness of speaking to the partner.**

Keywords-Avatar; face tracking; video chat; videoconferencing; teleconferencing; distant communication; telepresence

I. INTRODUCTION

Thanks to the maturity of motion tracking technologies, it became easy and inexpensive to use real-time avatars that are computer graphics animation reflecting the user's facial and body movements in real time. The price of hardware devices for motion tracking is rapidly decreasing. And even standard webcams enable robust face tracking. So, the cost of a tracking system is no longer a barrier to using real-time avatars. The labor to prepare a tracking system is also no longer the barrier. Users do not need to wear optical markers, and just need to stand or be seated in front of a camera. Thus, in terms of cost and easiness, there is almost no difference between using videos and real-time avatars when talking with remote conversation partners. Real-time avatars are now becoming considered an alternative to videos [1,2].

The most prominent advantage of real-time avatars over videos is the capability of transmitting the user's motion to remote sites without disclosing the user's appearance. A consciousness of being remotely watched is a well-known problem of video chat [3]. In avatar chat that is voice chat accompanied by real-time avatars, this consciousness is drastically mitigated since the images that capture the user's face and clothes are not shown to remote conversation partners. Without those information the users are still able to use a visual communication channel to exchange facial expressions and gestures. Here the question is whether pure human motion without any information of appearance is useful in distant communication.

In general, the addition of a video connection has been regarded as a single factor in past research [1,4-7]. However, that can actually be divided into two factors, i.e., the motion and appearance factors. A real-time avatar corresponds to the motion factor and does not include the appearance factor. To prove that real-time avatars are useful, it is necessary to demonstrate that the motion factor improves distant communication independently from the appearance factor. Therefore, we conducted an experiment to examine the two factors. To analyze the effects of the two factors separately, we prepared a visual representation that corresponds to the appearance factor and does not include the motion factor. As shown in Figure 1, we selected the photo of a remote conversation partner as that representation. Since many users of instant messengers put their photos in the buddy list, watching the partner's photo in voice chat is a popular situation.

II. RELATED WORK

This paper presents the result of comparing avatar chat, video chat, photo chat and voice chat. The superiority of photo chat or avatar chat over voice chat has been unclear, and even the superiority of video chat over voice chat has not been yet clarified sufficiently, as described below.

There are studies that observed the effects of photos on audio communication [8,9]. However, it is still unclear how presenting the still picture of a remote conversation partner to the speaker influences speech. There are also studies that observed the effects of real-time avatars on audio communication [1,10-13]. And it has not been yet demonstrated that the capability of transmitting human motion leads to improvements in distant communication.

Intuitively, videoconferences seem to be more useful than audio-only teleconferences [14]. But, there is a long history of discussing the necessity of sharing the live video of remote participants in teleconferencing, and there has been a consistent trend to deny the advantage of sharing the video [15,16]. To clarify the superiority of videoconferences over audio-only

978-1-4673-6404-1/13/$31.00 ©2013 IEEE

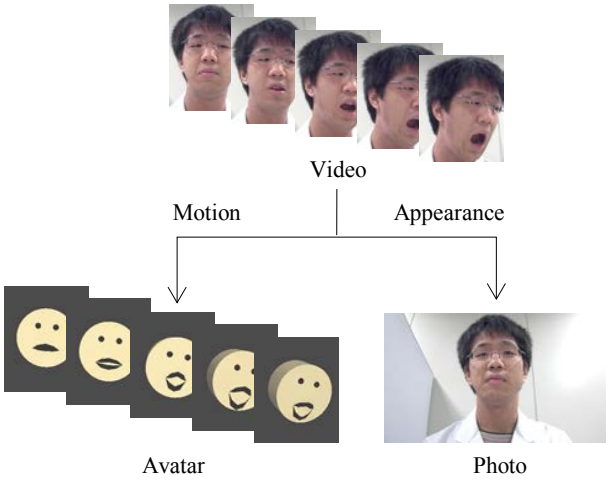


Figure 1. Motion and appearance in video.

communication, many studies tried to find positive effects produced by the addition of video connections. Some of them showed that video connections could induce positive responses from subjects [5,6,17], and could facilitate multiparty conversations [5]. In multiparty conversations, the addition of video connections clarifies who is talking to whom. However, in dyadic conversations that are one-to-one conversations between two persons, the speaker and the listener are naturally clear. Thus, there is no well-known positive effect in objective measures, e.g., conversational structure [4,5,17] and task performance [4,6,7], in dyadic conversations.

According to the previous studies described above, it seems to be harder to find a positive influence of additional communication channels in dyadic conversations than multiparty conversations. And it seems to be harder to do that in objective measures than subjective measures. Our study focused on the hardest condition. Namely, we tried to find objectively measured values that prove usefulness of photos, avatars, and videos in dyadic conversations. No previous study found such values.

The difference of our study from the previous studies is that we focused on more fine-grained structure of conversations, i.e., pauses in speech as shown in Figure 2, in which each box represents an utterance and each space between boxes represents a pause. As the figure shows, fewer pauses indicate smooth speaking and more pauses indicate awkward speaking. It was reported that the frequency of pauses is an indicator of the level of anxiety [18,19]. On the assumption that increased degree of smoothness of speech was an improvement in communication, we analyzed a single turn instead of analyzing turn-taking, which is a typical conversational structure that has been analyzed in many studies [17].

III. EXPERIMENT

A. Hypothesis

We conducted an experiment to prove that real-time avatars improve distant communication. Real-time avatars transmit the

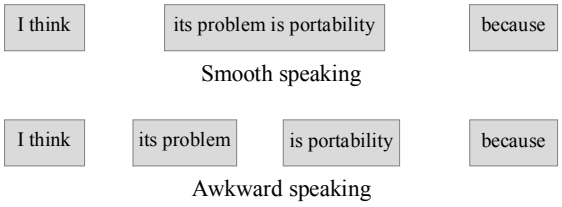


Figure 2. Smoothness of speech.

user's motion to remote sites without disclosing the user's appearance. Therefore, we expected that human motion delivered by real-time avatars improves distant communication and a lack of appearance does not affect that improvement. To clarify the relationship between the motion and appearance, we compared four chat styles of a 2x2 design: video chat (motion and appearance), avatar chat (motion but no appearance), photo chat (appearance but no motion), and voice chat (no motion, no appearance). We predicted that avatar chat and video chat, which deliver human motion, are better than voice chat. Since we assumed that the improvement would appear as a smooth speech, the hypotheses of the experiment were the followings.

Hypothesis 1: Compared with audio-only communication, a speech that is directed to the remote conversation partner is smoothened when the speaker can see the partner's video.

Hypothesis 2: Compared with audio-only communication, a speech that is directed to the remote conversation partner is smoothened when the speaker can see the partner's avatar.

B. Conditions

To examine the hypotheses, we prepared the four conditions of the 2x2 design shown in Figure 3. In all the conditions, the subject was seated at a desk on which there was a teleconferencing terminal, which consisted of a microphone speaker, a display, the camera for live video, and the camera for face tracking. The display was a 10-inch wide-screen LCD, which was much smaller than the displays that are usually used for videoconferences so that we could confirm that even a small-size visual representation of a remote conversation partner improves communication. The details of each condition are described below.

Voice condition (no motion, no appearance): This condition was identical with a normal voice chat. The subject talked to the remote conversation partner through only the microphone speaker. To let the subjects intuitively recognize that they shared no visual information with the partner, the display was blacked out and the two cameras were covered with a box.

Photo condition (appearance but no motion): This condition was similar to a voice chat in instant messengers. The difference between this condition and the voice condition was the existence of the remote conversation partner's facial photo shown on the display. A small photo of the subject was also shown at the display's bottom right corner so that the subjects could know how they were shown on the remote display. This self-portrait photo was taken at the beginning of the experiment. In the same manner as the voice condition, the two

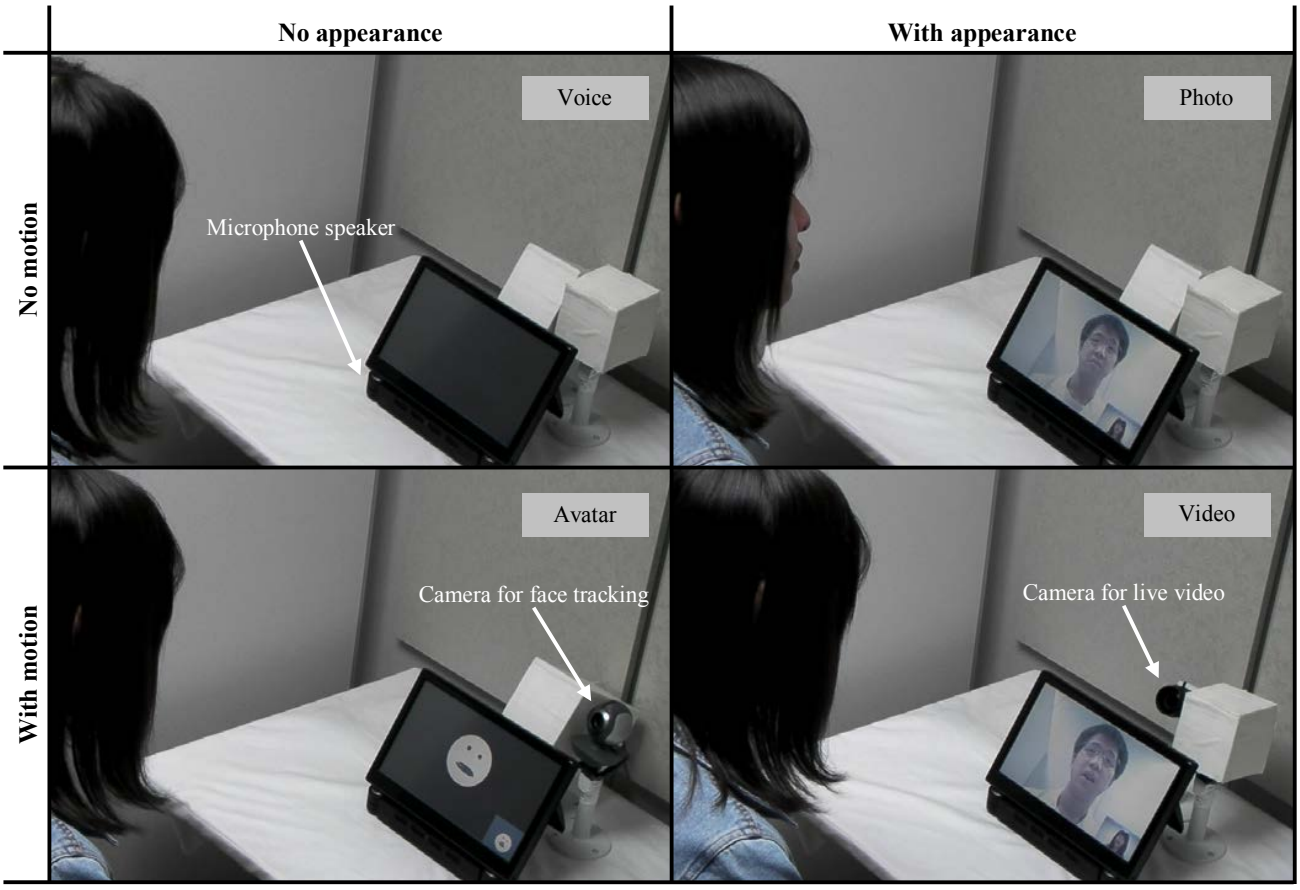


Figure 3. Conditions of the experiment.

cameras were covered with a box and the conversation occurred only through an audio connection.

Avatar condition (motion but no appearance): In this condition the display showed an anonymous avatar, which reflected the remote conversation partner's head and lip motions but did not reflect anything of the partner's facial image, as shown in Figure 1. The camera for live video was covered but the camera for face tracking was uncovered. The three-dimensional model of the avatar consisted of a cylindrical head, lips, and eyeballs. The shape of the head was not a sphere but a cylinder, since a cylindrical head is better to show changes in the facial direction than a spherical head. The size of the head was almost equal to the size of the image of the partner's head in the photo and video conditions. The color of the lips was not a red but a dark gray, since we used the minimum number of colors to draw the avatar in order to make the design of the avatar as simple as possible. We used light yellow to draw the head, and used dark gray to draw the lips, the eyeballs, and the background. The head translated and rotated with six degrees of freedom. Each lip was transformed based on the three-dimensional positions of eight markers. The eyeballs were tiny spheres attached to the fixed positions of the face. The head and lips moved at thirty frames per second according to the sensor data sent from the face tracking software (faceAPI), which was running in the remote terminal and capturing the partner's movements. The subject's movements were also captured, and their avatar was shown on the remote display. A small avatar of the subject was shown at the bottom right corner of the display so that the subjects could know how they were shown on the remote display and could also confirm how precisely the avatar was able to reflect human movements.

Video condition (motion and appearance): This condition was identical to a normal video chat. Only the camera for live video was uncovered. The display showed the live video of the remote conversation partner's face, and also showed a small mirror window at the bottom right corner. The resolution of the partner's video was 640 pixels by 375 pixels, and its frame rate was thirty frames per second.

C. Subjects

This experiment was within-subject design, so each subject experienced all of the above four conditions. The order of experiencing the conditions was counterbalanced to make the influence of order effects as small as possible. To include all kinds of the order of the four conditions, the number of the collected subjects was twenty four that is equal to four

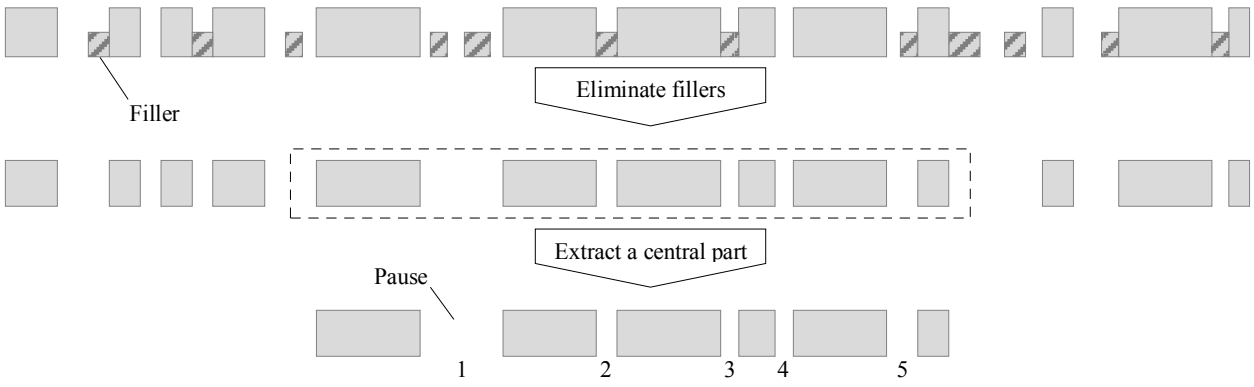


Figure 4. Method to count the number of pauses.

factorial. Twenty-four undergraduate students who lived near our university campus participated in the experiment.

D. Task

To measure the frequency of pauses stably we had to record long speech. Simultaneously, we also had 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. Thus, we elaborated a task in which the subject could continue to talk for more than one minute without the partner's interference. At the beginning of each condition, the subject was asked by the conversation partner to talk about the pros and cons of a certain gadget and possible improvements to that gadget.

In all the conditions, the same experimenter played the role of the conversation partner. 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.

Before experiencing the four conditions, the subjects practiced the task in a face-to-face environment. In this FTF session the subject was familiarized with the task and the experimenter.

Since each of the subjects experienced the first FTF session and the four conditions, we prepared five gadgets as conversational topics, i.e., e-book readers, handheld game consoles, smartphones, robotic vacuum cleaners, and 3D TVs. The order of these gadgets to be talked about was counterbalanced. The experimenter told the subject which gadget to talk about right when the conversation began. We did not disclose the next topic beforehand.

We did not ask the subject to talk for more than a certain duration, so the subject could stop talking anytime. However, since the five gadgets are attracting considerable attention recently, almost all of the subjects knew the pros and cons 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 that was enough to observe the difference in the frequency of pauses.

E. Data Collection

To measure the smoothness of the subjects' speech, we counted the number of pauses in the recorded speech as shown in Figure 4. First, we eliminated all pause fillers from the speech. To determine what kind of utterance was regarded as a pause filler, we referred to research papers that presented a list of filler words [20]. A pause that included fillers within the pause was replaced with a single pause, because the fillers within a pause divide an actual single pause into multiple pauses and inflate the number of pauses. Also, a filler that concatenated two utterances was replaced with a pause, because a filler between utterances actually breaks the speech like a silent pause. A filler that just preceded or followed an utterance did not change the number of pauses, but we also eliminated that kind of a filler.

Next, we extracted the central part of the speech. The speech of every subject 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 smoothness 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 of the subjects visited a store to buy one of the gadgets on the day before the experiment. Further, the ending part tended to be needlessly awkward if the subject made an extra effort to continue talking.

Finally, we counted the number of pauses included in the central part. 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.

How the recorded speeches were processed is as follows. We used a multimedia annotation tool (ELAN) to transcribe them. We entered the beginning time, ending time, and

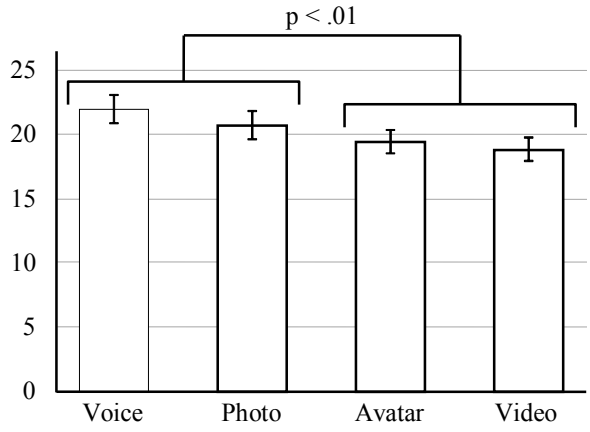


Figure 5. Average frequency of pauses.

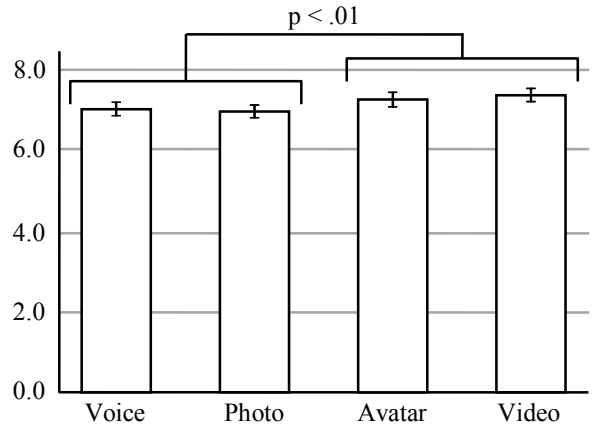


Figure 6. Average speech rate.

transcript of all the utterances. In this data entry, we distinguished pause fillers from normal utterances. After we finished the data entry, the data were exported as a text file and pasted into a spreadsheet file. We used a spreadsheet software to count the number of syllables, extract a central part, and count the number of pauses.

IV. Result

We compared the four conditions to find the effects of the motion and appearance factors. Since each factor had two levels as shown in Figure 3 and each subject produced the data of all conditions, we conducted 2x2 two-way repeated-measures ANOVA. A capability to talk smoothly varies with the individual. Thus, to compare the four conditions within each subject, we used repeated-measures ANOVA instead of simple ANOVA. All the results described below are the results of analyzing the central two-hundred syllables of each speech.

In Figure 5 and 6, each box represents the mean value and each bar represents the standard error of the mean value. Figure 5 shows the result of comparing the frequency of pauses, which was actually the number of pauses included in the extracted part of each speech. We found a strong main effect of the motion factor ($F(1,23)=13.307$, $p<.01$) in the frequency of pauses. Figure 6 shows the result of comparing speech rate, which was the number of syllables per second. We also found a strong main effect of the motion factor ($F(1,23)=19.017$, $p<.01$) in the speech rate. In these two analyses, the main effect of the appearance factor and the interaction between the two factors were not significant. These results meant that the visually presented motion of the remote conversation partner reduced the frequency of pauses and increased the speed of speaking to the partner. And the absence of the partner's appearance did not significantly affect these influences.

The above results prove the two hypotheses that the presentation of videos and avatars smoothens distant audio communication. The subjects' speech that was directed to the remote conversation partner was smoothened by the presentation of a live video of the partner. The speech was also smoothened by the presentation of an anonymous avatar that did not show the partner's facial image and showed only the partner's motion of a head and lips. There was no clear difference between the video and avatar in the degree of smoothening the speech, and there was no clear effect of the partner's photo on the speech. The effects of presenting the partner's facial image to the subjects were invisible.

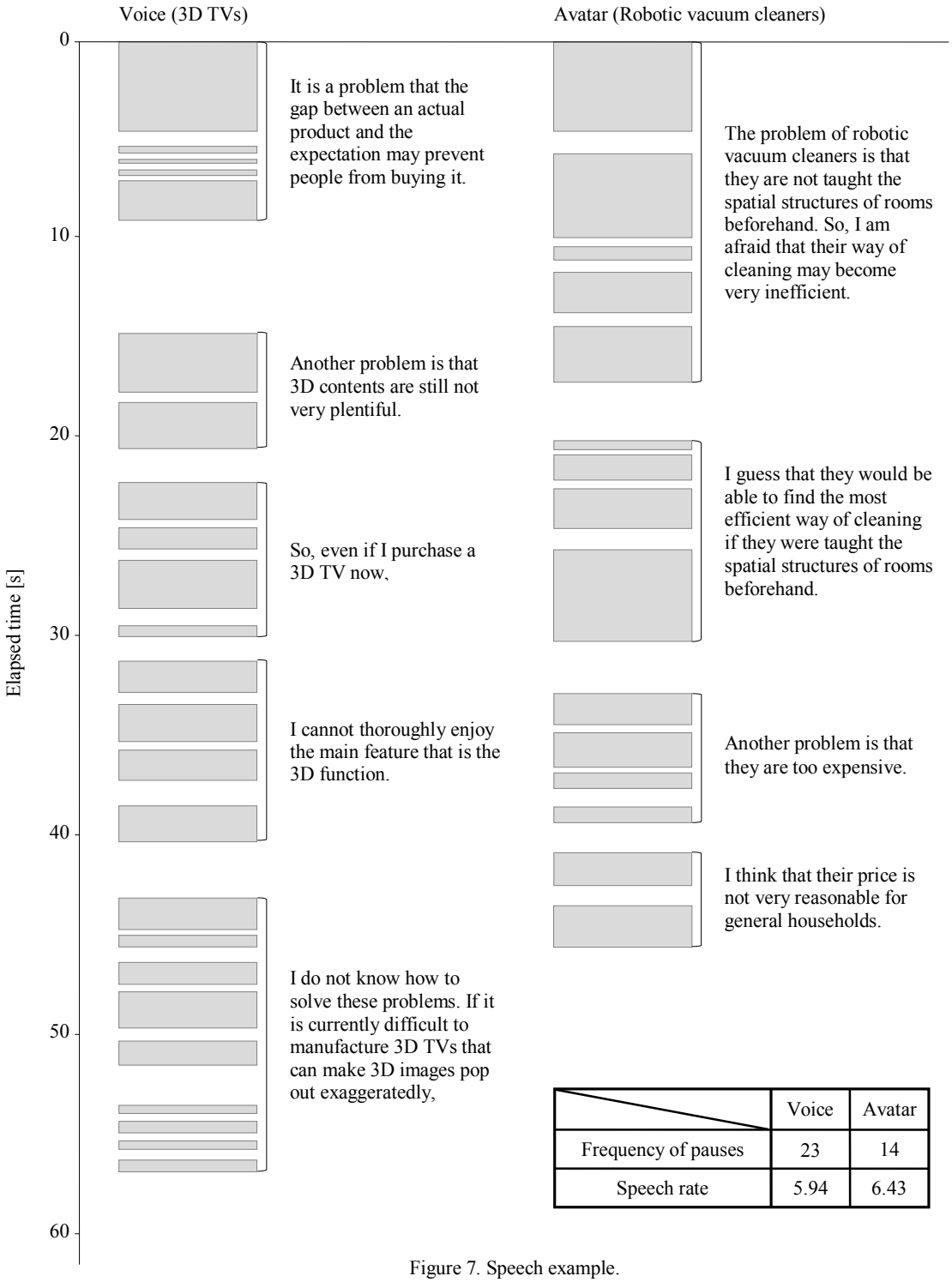
In Figure 7, you can see the difference between the awkward and smooth speech of one of the subjects. The figure arranges the pattern of pauses and the transcript in the central two-hundred syllables of the two speeches. The vertical axis is the elapsed time. In the pause patterns each box represents an utterance and each space between boxes represents a pause. The left side is the voice condition in which the topic was 3D TVs, and the right side is the avatar condition in which the topic was robotic vacuum cleaners. You can see that the left awkward speech included more pauses and took longer time.

We also analyzed the following values: the time of the pauses, the ratio of the time of the pauses to the time of the speech, the number of the fillers, the time of the fillers, the ratio of the time of the fillers to the time of the speech, and the ratio of the time of the fillers to the time of the pauses. In all of these analyses, the main effects of the motion and appearance factors and the interaction between the two factors were not significant.

V. Discussion

The results of the experiment showed that real-time avatars and live videos are equally effective for audio communication as an additional channel. The common capability of the avatars and videos – presenting a remote person's motion – improved distant communication. And the videos' capability that was not owned by the avatars – presenting a remote person's appearance – did not lead to a significant improvement. Hence, the motion seems to be enough to smooth distant communication, and the appearance seems to be redundant information for the smoothness.

The presentation of a remote conversation partner's face could not produce significant effects, even though we degraded the audio and avatar conditions to strengthen the appearance factor as follows. The design of the avatar was made very



	Voice	Avatar
Frequency of pauses	23	14
Speech rate	5.94	6.43

Figure 7. Speech example.

simplistic and a little weird. In the voice condition the display was blanked out instead of showing the still image of the avatar.

The important thing to understand the results is that past studies have not found any objectively measured difference between video-mediated and voice-mediated dyadic conversations. Our study found such differences, which were reduction in the frequency of pauses and increase in the speed of speech. And our study also found that even real-time avatars are able to produce the same differences. Live videos may have another observational effect on dyadic conversations, and real-

time avatars may not be able to produce that effect. But, such an effect has not been found so far.

We interviewed the subjects after they experienced all conditions. The results of the interviews show that live videos tend to invite stronger criticism than real-time avatars. Several subjects criticized the video condition because they were conscious of being watched by a remote conversation partner, which is a well-known problem of video connections [3]. Other criticisms were also caused by well-known problems: the lack of eye contact and the delay of the video transmission. It was interesting that the avatar condition did not suffer any of these criticisms. Even though in the avatar condition the subjects' motion was always watched by the partner, eye contact was not established because of the avatar's fixed eyeballs, and the delay of the avatar's animation was actually twice as long as that of the video transmission – the avatar's delay being 0.3 second and the video's delay being 0.15 second. Of course, the avatar was not totally free from criticism. For example, the avatar was criticized for its unnatural visual design.

Compared with live videos, real-time avatars have the advantage of small-size data. The parameters necessary to draw our avatar were the six degrees of freedom of the head and the eight markers of each lip. So, the total amount of the data was sixty-six floating-point numbers. On the other hand, the amount of the data necessary to draw a frame of the live video we used in the experiment was 240,000 pixels. Assuming that the size of the data of a pixel was equal to that of a floating-point number, the amount of network bandwidth consumed by live videos is more than thirty-six hundreds times as large as that of real-time avatars.

As described above, real-time avatars are a promising substitute for live videos in distant communication. Of course, there are many situations where appearance information delivered by live videos is required. For example, you may like to see a conversation partner's face if the partner is your acquaintance. Visual information is often essential in telemedicine and distance learning. So, we have no intention to argue that the appearance is redundant information in all cases.

In this study we used only one third of the surface of a 10-inch wide-screen LCD to show the remote conversation partner's video and avatar. So the size of the video and avatar was small enough to be displayed on a tablet PC or a large-size smartphone. It is interesting to investigate how small videos or avatars can smoothen audio communication. Video calling on mobile phones may have the same effect.

It is known that video connections enhance social presence [5]. Clarifying the relationship between social presence and the smoothness of speech is a topic for future study. There are several techniques to enhance social presence in videoconferences [21]. It is interesting to see how these techniques reduce the frequency of pauses and increase the speed of speech.

In this study, face tracking technologies produced the avatar's motion from the remote conversation partner's motion. However, a looser coupling between the avatar and the partner may be effective. Past research has developed technologies that convert the vocal data of speech into an avatar's motion of speaking and also listening [22]. The motion generated by those technologies may be able to improve distant communication. If people are not sensitive to the fidelity of motion, arbitrarily created motion is likely to be effective. It was reported that there was no significant difference in the degree of social presence between a human-controlled avatar and an automatically moving agent [23]. Examining how much fidelity of motion is necessary for the avatar to be effective is a future study.

The real-time avatar we used in this study was a two-dimensional computer graphics animation. A physical embodiment may enhance the capability to smoothen speeches. It seems to be interesting to use an anonymous embodied robotic avatar that has a neutral face without a specific age and gender instead of our anonymous two-dimensional avatar [24]. It also seems to be interesting to test another way of physical embodiment that is movable displays [25,26]. They can emphasize the motion delivered by a live video or a real-time avatar.

This study dealt only with head and lip motions. There is a possibility that other motions are also be able to improve distant communication, e.g., eyebrows, hands, and positional movement. Our avatar did not have eyebrows and hands. This indicates that facial expression and gesture are not essential to the increase in the degree of smoothness of speech. However, they may enhance the capability to smoothen speeches. There are several previous studies that equipped audio communication channels with a function to share positional movements in a physical space [27,28]. It is interesting to see how shared positional movements change the smoothness of speech.

VI. Conclusion

This paper showed how distant communication is improved by real-time avatars that are computer graphics animation reflecting the user's facial and body movements in real time. We conducted an experiment to observe differences between video chat and voice chat, and examined whether the same differences could be produced by avatar chat that was voice chat accompanied by real-time avatars. As a result, we found that the degree of the smoothness of speech differed between video chat and voice chat and the degree increased in both of video chat and avatar chat.

In this study we found that the existence of video connections reduces the frequency of pauses in a speech and increases the speed of speaking. Traditionally, video connections have been considered unimportant to distant communication. Past studies have not found any objectively measured difference between video-mediated and voice-mediated dyadic conversations, since it is more difficult to find significant effects of video connections in dyadic conversations than multiparty conversations, and in objective measures than subjective measures. We used objective measures and found the significant effects in dyadic conversations.

Our experiment also demonstrated that even avatars, which show the motion of a remote conversation partner and hide the appearance of the partner, are able to reduce the frequency of pauses and increase the speed of speech. The absence of the

partner's appearance did not significantly affect these influences. In terms of the influences, videos include redundant information that is the appearance. Our avatar moved only its head and lips, and did not have eyebrows and hands. This simple avatar and a live video had the same effects on audio communication.

We conclude that real-time avatars are a promising substitute for live videos in distant communication since both have the same objectively measured effects on dyadic conversations. In our experiment live videos tended to suffer various criticisms. So, we think that real-time avatars are a better choice unless the conversation participants have some special reason to share their appearance.

ACKNOWLEDGMENT

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] S. Junuzovic, K. Inkpen, J. Tang, M. Sedlins, and K. Fisher, "To see or not to see: a study comparing four-way avatar, video, and audio conferencing for work," *Proc. GROUP 2012*, 2012, pp. 31-34.

[2] J. C. Tang, C. Wei, and R. Kawal, "Social Telepresence Bakeoff: Skype Group Video Calling, Google+ Hangouts, and Microsoft Avatar Kinect," *Proc. of CSCW 2012 (Companion)*, 2012, pp. 37-40.

[3] E. Bradner, and G. Mark, "Social Presence with Video and Application Sharing," *Proc. GROUP 2001*, 2001, pp. 154-161.

[4] A. H. Anderson, A. Newlands, J. Mullin, A. Fleming, G. Doherty-Sneddon, and J. M. Van Der Velden, "Impact of Video-Mediated Communication on Simulated Service Encounters," *Interacting with Computers*, vol. 8, no. 2, 1996, pp. 193-206.

[5] O. Daly-Jones, A. F. Monk, and L. Watts, "Some Advantages of Video Conferencing over High-quality Audio Conferencing: Fluency and Awareness of Attentional Focus," *International Journal of Human-computer Studies*, vol. 49, no. 1, 1998, pp. 21-58.

[6] J. S. Olson, G. M. Olson, and D. K. Meader, "What Mix of Video and Audio Is Useful for Small Groups Doing Remote Real-time Design Work?" *Proc. CHI 95*, 1995, pp. 362-368.

[7] G. P. Radford, B. F. Morganstern, C. W. McMickle, and J. K. Lehr, "The Impact of Four Conferencing Formats on The Efficiency and Quality of Small Group Decision Making in a Laboratory Experiment Setting," *Telematics and Informatics*, vol. 11, no. 2, 1994, pp. 97-109.

[8] R. A. Colburn, M. F. Cohen, S. M. Drucker, S. L. Tiernan, and A. Gupta, "Graphical Enhancements for Voice Only Conference Calls," *Microsoft Research Technical Report, MSR-TR-2001-95*, 2001.

[9] M. Tanis, and T. Postmes, "Two Faces of Anonymity: Paradoxical Effects of Cues to Identity in CMC," *Computers in Human Behavior*, vol. 23, no. 2, 2007, pp. 955-970.

[10] J. N. Bailenson, N. Yee, D. Merget, and R. Schroeder, "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*, vol. 15, no. 4, 2006, pp. 359-372.

[11] G. Bente, S. Ruggenberg, N. C. Kramer, and F. Eschenburg, "Avatar-Mediated Networking: Increasing Social Presence and Interpersonal Trust in Net-Based Collaborations," *Human Communication Research*, vol. 34, no. 2, 2008, pp. 287-318.

[12] M. Garau, M. Slater, S. Bee, and M. A. Sasse, "The Impact of Eye Gaze on Communication Using Humanoid Avatars," *Proc. CHI 2001*, 2001, pp. 309-316.

[13] S. Kang, J. H. Watt, and S. K. Ala, "Communicators' Perceptions of Social Presence as a Function of Avatar Realism in Small Display Mobile Communication Devices," *Proc. HICSS 2008*, 2008.

[14] E. A. Isaacs, and J. C. Tang, "What Video Can and Can't Do for Collaboration: a Case Study," *Multimedia Systems*, vol. 2, no. 2, 1994, pp. 63-73.

[15] E. A. Isaacs, and J. C. Tang, "What Video Can and Can't Do for Collaboration: a Case Study," *Multimedia Systems*, vol. 2, no. 2, 1994, pp. 63-73.

[16] R. Pye, and E. Williams, "Teleconferencing: Is Video Valuable or Is Audio Adequate?" *Telecommunications Policy*, vol. 1, no. 3, 1977, pp. 230-241.

[17] A. J. Sellen, "Remote Conversations: The Effects of Mediating Talk with Technology," *Human-Computer Interaction*, vol. 10, no. 4, 1995, pp. 401-444.

[18] A. M. Goberman, S. Hughes, and T. Haydock, "Acoustic characteristics of public speaking: Anxiety and practice effects," *Journal of Speech Communication*, vol. 53, no. 6, 2011, pp. 867-876.

[19] J. A. Harrigan, I. Suarez, and J. S. Hartman, "Effect of Speech Errors on Observers' Judgments of Anxious and Defensive Individuals," *Journal of Research in Personality*. vol. 28, no. 4, 1994, pp. 505-529.

[20] S. Ishihara, and Y. Kinoshita, "Filler Words as a Speaker Classification Feature," *Proc. SST 2010*, 2010, pp. 34-37.

[21] A. Prussog, L. Muhlbach, and M. Bocker, "Telepresence in Videocommunications," *Proc. Annual Meeting of Human Factors and Ergonomics Society*, 1994, pp. 25-38.

[22] H. Ogawa, and T. Watanabe, "InterRobot: Speech-Driven Embodied Interaction Robot," *Advanced Robotics*, vol. 15, no. 3, 2001, pp. 371-377.

[23] A. M. von der Putten, N. C. Kramer, and J. Gratch, "Who's there? Can a Virtual Agent Really Elicit Social Presence?" *Proc. PRESENCE 2009*, 2009.

[24] K. Ogawa, S. Nishio, K. Koda, G. Balistreri, T. Watanabe, and H. Ishiguro, "Exploring the Natural Reaction of Young and Aged Person with Telenoid in a Real World," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 15, no. 5, 2011, pp. 592-597.

[25] H. Nakanishi, K. Kato, and H. Ishiguro, "Zoom Cameras and Movable Displays Enhance Social Telepresence," *Proc. CHI 2011*, 2011, pp. 63-72.

[26] D. Sirkin, and W. Ju, "Consistency in Physical and On-screen Action Improves Perceptions of Telepresence Robots," *Proc. HRI 2012*, 2012, pp. 57-64.

[27] M. Flintham, R. Anastasi, S. Benford, T. Hemmings, A. Crabtree, C. Greenhalgh, T. Rodden, N. Tandavanitj, M. Adams, and Ju. Row-Farr, "Where On-Line Meets On-The-Streets: Experiences With Mobile Mixed Reality Games," *Proc. CHI 2003*, 2003, pp. 569-576.

[28] H. Nakanishi, S. Koizumi, T. Ishida, and H. Ito, "Transcendent Communication: Location-Based Guidance for Large-Scale Public Spaces," *Proc. CHI 2004*, 2004, pp. 655-662.