To See or Not to See: A Study Comparing Four-way Avatar, Video, and Audio Conferencing for Work

Sasa Junuzovic, Kori Inkpen, John Tang
Microsoft Research
Redmond, WA
{sasajun,kori,johntang}@microsoft.com

Mara Sedlins
Dept. of Psychology, UW
Seattle, WA
sedlins@uw.edu

Kristie Fisher
Microsoft
Redmond, WA
kfisher@microsoft.com

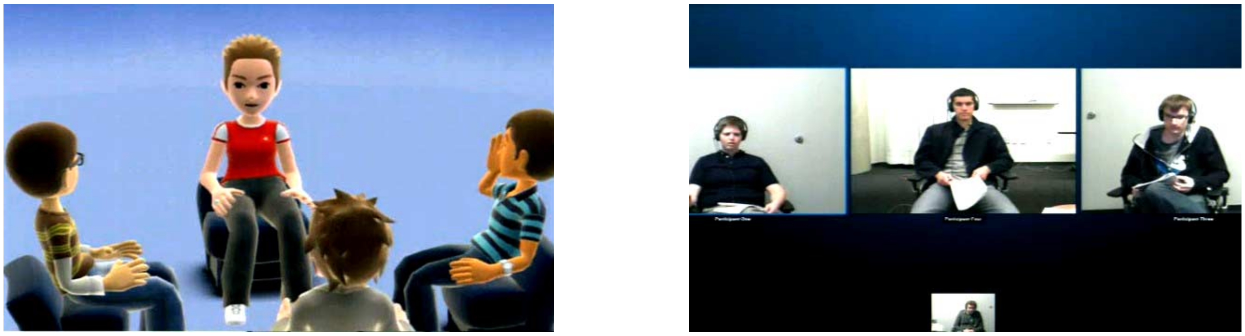

Figure 1. Four-way Microsoft Avatar Kinect (left) and Skype Group Video (right) conference

ABSTRACT

We conducted a study comparing avatar conferencing with video and audio conferencing for work scenarios. We studied nine four-person teams using a within-subjects design that measured users' perceptions and preferences across the conferencing conditions. Video was rated highest in all measures. Avatar and Audio were rated similarly, except for sociability, where Avatar was rated higher than Audio, and realism, where Avatar was rated lower than Audio. While users appreciated how avatar conferencing brought them together in a common virtual space, they found the cartoon avatars to be inappropriate for a professional discussion. As a result, participants preferred Video the most and Avatar the least for a business meeting. Lower ratings for the avatar condition were partly due to users' frustrations when the avatar system did not track them perfectly. When assuming a "perfect" system, preference for Avatar increased significantly while preference for Audio and Video remained unchanged.

Categories and Subject Descriptors

H5.3. Group and Organization Interfaces (CSCW).

Keywords

Conferencing; audio; video; avatar; distributed teams.

1. INTRODUCTION

Recently available commercial technologies have enabled new forms of synchronous conferencing using video and avatars. Desktop video services enable n-way video conferences where each participant can be seen at all times. Meanwhile, avatar representations in a virtual world, popularized by online gaming, afford meeting and interacting together in a virtually constructed setting. As these technologies gain popularity in consumer markets, we expect they will also be used in workplace meetings. However, the visual representation of meeting participants may affect the interactions that occur in these mediated environments, and we explore their impact in this study.

We were interested to see how avatar conferencing compared with more traditional audio and video conferencing in the workplace. Avatar conferencing enables users to collaborate together through virtual avatars that represent each person's bodily movements in both gaming [6] and commercial applications [3]. Until recently, embodying these avatars required users to make manual keyboard and mouse commands that translated into basic avatar actions (e.g., walk forward, wave). More recent systems, such as Microsoft's Avatar Kinect, capture users' motions through depth cameras and use the depth information to animate 3D cartoon avatars to reflect users' body movements. This natural user interface can automatically convey non-verbal cues.

Compared to video conferencing, avatar conferencing has many potential user experience advantages. Virtual avatars abstract away the users' real environments, which can mitigate privacy concerns that video evokes [2] and also enable users to manage their appearance (for example, looking professional when joining from home). Furthermore, virtual worlds can create a common meeting space for distributed team members connecting from diverse settings. Virtual avatars can even synthesize non-verbal cues, such as turning toward the current speaker, which could strengthen the sense of presence (defined as the feeling of being socially present with people at a remote location [8]).

The goal of our study was to compare multi-party avatar conferencing to video and audio conferencing for workplace collaboration. Several prior studies have already compared video

GROUP'12, October 27–31, 2012, Sanibel Island, Florida, USA.
Copyright 2012 ACM 978-1-4503-1486-2/12/10...$15.00.

and audio conferencing and documented advantages of the visual channel [10], [11]. In particular, social presence increased with the bandwidth of the communication medium (e.g., social presence was higher for video than audio) [9]. We were interested in how avatar conferencing would compare with audio and video.

Previous work by Bente et al. [1] compared avatar conferencing with audio, video, and text conferencing in two-way conferences between strangers. They found that avatar and video conferencing were similar with respect to user satisfaction, trust, and social presence. However, it is unclear how these results generalize to multi-party groups, where the fidelity of visual cues, such as facial expressions and gaze awareness, become more important. Prior work [7], [10] has shown that using video to provide these visual cues can help reduce potential interaction ambiguities that can occur with more than two people in a conference. Thus, it is important to re-evaluate avatar, audio, and video conferencing in multi-party scenarios. We chose to examine four-way conferencing as a large enough group to exercise both gaze awareness and non-verbal communication cues.

By studying avatar, video, and audio conferencing, we could see how adding different representations of visual cues to the shared audio communication (which was common across all conditions) affected the collaboration. Intuitively, adding avatar visual cues should improve the experience over audio alone. However, avatars are not as high fidelity as video. Moreover, cartoon aspects of avatars may conflict with users' expectations of how remote people should be presented visually. Thus, it is interesting to explore how the avatar experience compares to video.

2. METHODOLOGY

Our user study compared audio-only (*Audio*) and audio-video (*Video*) conferencing with 3D-avatar conferencing (*Avatar*) using existing commercial tools. In all three conditions, Skype group audio conferencing provided the audio channel. For Video conferencing, we used Skype Group Video Calling, configured to show all of the remote participants aligned horizontally, as shown in Figure 1 (right panel).

For Avatar conferencing, we used a beta version of Microsoft Avatar Kinect, an XBOX 360 avatar chat application. Avatar Kinect animates a cartoon avatar in a virtual world based on a user's movements in the real world. The Kinect sensor tracks the user's upper body motion, including torso and arm positions, as well as some facial features, namely, mouth and eyebrows. Based on the tracking data, the avatar mimics posture changes, hand waves, head turns, lip movements, and facial expressions in real time. However, unlike video, these visual cues are presented by animating a computer-generated, cartoon avatar. We chose a virtual world where avatars sat in a circle (see Figure 1, left panel). Each participant had a third-person view of the world as if standing behind their own avatar. This view most closely matched the view of others in Skype, although Skype's preview of oneself is frontal and not over-the-shoulder. In Figure 1, the local user, whose avatar is at the bottom of the screen, can see that all avatars are looking at the avatar at the top of the screen, thus conveying a shared sense of gaze awareness.

2.1 Participants and Procedure

We recruited 36 participants (16 females, 20 males), in 9 groups with 4 participants per group, from within Microsoft, a large software company. As prior work has demonstrated, participants' familiarity with each other affects their conferencing experience [4], and since participants in a business meeting are usually familiar with each other, we recruited participants who already knew each other. Our participants will be referred to as ($P_{x,y}$) where x is the group number and y is the participant number within that group.

Each group of participants used all three conferencing technologies and worked through three brainstorming meetings that were equivalent in structure but involved different, although related, topics. We chose a brainstorming task since it is a common business practice that requires participation from everyone and may involve persuasion and negotiation, for which visual cues are important. We selected discussion topics (features of next generation mobile devices) that are important to our participants' company to provide some inherent motivation for the task. One discussion focused on smartphones, another on tablet devices, and the remaining discussion focused on mobile search. For each condition, the participants brainstormed for about 10 minutes. The study administrators then interrupted them and asked them to agree on the top four features discussed during the brainstorming and their priority order.

Each of the four conference participants was placed in a different room that had a 40" LCD 1080p TV, a headset, an XBOX 360, and a computer. For the Avatar condition, participants spent 5-10 minutes creating avatars that looked like them by tailoring avatar attributes, such as hair style, and facial features. For all three conferencing technologies, the participants used a headset to hear and talk to each other. In the Audio condition, the TV showed a blank computer desktop. In the Video condition, participants saw video windows of remote participants on the TV through Skype Group Video Calling, as shown in Figure 1 (right). In the Avatar condition, participants saw their avatars in a virtual location together with the avatars of the remote participants through Microsoft Avatar Kinect, as shown in Figure 1 (left).

2.2 Experimental Design

We used a within-subjects study design with condition order counterbalanced using a Partial Latin Square design. All groups performed the brainstorming tasks in the same order, starting with smartphone features, followed by tablet features, and finishing with mobile search features on smartphone and tablet devices. We chose a within-subjects design to reduce the impact of individual differences and enable users to compare across conditions.

At the start of a study session, all participants completed an initial questionnaire that asked for their demographic information and prior experience with smartphones, tablets, and avatars. The participants also completed a questionnaire after each brainstorming task. Finally, all participants completed a questionnaire at the end of the session that compared across conditions and took part in a group debriefing session.

The post-task questionnaires consisted of different groups of questions: *social presence*; *conversation mechanics and non-verbal communication cues*; and *realism*. The social presence questions used anchored seven-point scales, while the remainder of the questions utilized seven-point Likert-type scales from 1 (strongly disagree) to 7 (strongly agree).

The *social presence* questions focused on how the users perceived the various conferencing technologies from a social perspective. These questions probed four dimensions that have been shown to differentiate social presence in telecommunications [9]: *impersonal-personal, cold-warm, insensitive-sensitive,* and

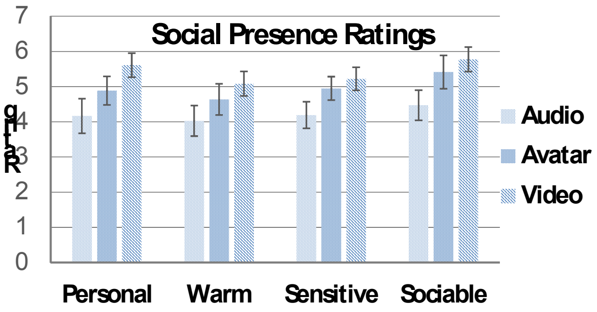

Figure 2. Average ratings for each social presence factor.

unsociable-sociable. The questions regarding *conversation mechanics* and *non-verbal cues* probed for the impact of the conferencing technologies on the conversation flow and non-verbal cue awareness. Sample questions included: "I could easily tell who was speaking" and "I could speak up in the discussion without interrupting someone else." The questions focusing on *realism* were included to help us understand how cartoon avatars compared to audio and video for communication purposes. Sample questions included: "It was just as though we were all in the same room", "The other people seemed real". Participants were also invited to provide free-form comments, reactions, likes, and dislikes about each system.

After completing all three tasks, participants were asked to rate each condition on a scale from 1 to 10, where 1 was "not useful at all" and 10 was "extremely useful." This question was repeated after asking the participants to assume perfect system performance, looking past current system flaws, such as poor motion tracking or audio lag.

3. RESULTS

Analyses were performed using Aligned Rank Transform, a new technique to enable use of parametric statistics on non-parametric data [12]. There were no significant effects of gender.

3.1 Social Presence

Significant main effects of condition were found for each of the social presence factors: impersonal-personal ($F_{2,70}$=16.33 p<.001), cold-warm ($F_{2,70}$=6.88 p<.005), insensitive-sensitive ($F_{2,70}$=19.07 p<.001), and unsociable-sociable ($F_{2,70}$=16.33 p<.001). Figure 2 shows the mean rating for each factor. Bonferroni corrected post-hoc pairwise comparison revealed that for all of the factors, the Video condition was rated significantly higher than Audio (*p*<.005). The Video condition was also rated significantly higher than the Avatar condition for the *personal* and *sensitive* dimensions (*p*<.005). However, no significant differences were found between Video and Avatar for *warm* (*p*=.19) and *sociable* (*p*=.47). The Avatar condition was rated significantly higher than Audio for being *sociable* (*p*<.005) but was not significantly different from Audio for the other three factors (*p*>.09). Participants' comments supported these results: "*Audio was the least personal and hardest to use in the realm of an unstructured meeting. Video was the most personal and the easiest to use. Kinect was easy to use and slightly more personal than audio* ($P_{4,3}$)". "*I liked the video the best because you can see how people react; get to know their personalities + it is like being in the same room as people. The avatar was a cool way at looking at conferences but I think it could be distracting if the discussion is business related. Audio is good, just very impersonal* ($P_{6,1}$)*.*"

Table 1. Results for non-verbal cues and realism. Significant differences were found for all questions (p<.001).

Non-Verbal Cues	
I could perceive and respond to non-verbal cues of people in the discussion	$F_{2,70}$=66.77
I could speak up in the discussion without interrupting someone else	$F_{2,70}$=26.76
I could speak up in the discussion without being interrupted by someone else	$F_{2,70}$=13.86
I got a good idea of how the other people were reacting	$F_{2,70}$=64.79
I got a real impression of personal contact with the other people	$F_{2,70}$=16.30
I could easily access the other people's reactions to what had been said	$F_{2,70}$=40.01
I could easily tell whom other people are directing their comments toward	$F_{2,70}$=16.39
Realism	
The discussion environment had a great sense of realism.	$F_{2,70}$=24.30
It was like having a face-to-face discussion.	$F_{2,70}$=35.70
It was just as though we were all in the same room.	$F_{2,70}$=23.82
The other people seemed "real."	$F_{2,70}$=34.80

3.2 Non-verbal Cues

The questionnaire also probed participants' impressions of non-verbal cues in the conferencing conditions, the seamlessness of the conversation, and whether people felt they could speak up without being interrupted (see Table 1). A significant main effect of condition was found for each question and Bonferroni corrected post-hoc pairwise comparison revealed Video was rated significantly higher than Avatar and Audio (*p*<.005), with no significant differences between the Avatar and Audio conditions (*p*>.07). Commenting on Video, one participant said: "*Much more able to glean subtle reactions* ($P_{10,1}$)."

3.3 Realism

Four questions probed the realism of the conferencing environments and how close it was to "having a face-to-face discussion" (see Table 1). Significant main effects of condition were found for each question. Bonferroni corrected post-hoc pairwise comparison revealed that for all of these questions Video was rated significantly higher than Avatar and Audio (*p*<.001) and no significant differences were found between Avatar and Audio (*p*>.05). When asked whether "it was just as though we were all in the same room," the Video condition was rated significantly higher than Avatar and Audio (*p*<.05), and the Avatar condition was rated significantly higher than the Audio condition (*p*<.005). As one participant commented, "*The people sitting in the one virtual room in the avatar space is a really important factor – makes you feel you are all at the same table and on equal footing* ($P_{5,4}$)".

3.4 Overall Preference

Consistent with the above ratings (on a scale from 1-10), a significant main effect of condition was found for overall preference ($F_{2,70}$=15.66 p<.001). Bonferroni corrected post-hoc pairwise comparison revealed that Video (m=8.9) was rated as significantly more useful for distributed meetings than Audio (m=6.9) and Avatar (m=5.1) (*p*<.005), with 78% of participants rating Video highest. While the Avatar condition was rated similarly to (and sometimes higher than) Audio on all of the individual dimensions examined, it was rated significantly lower than Audio in terms of overall usefulness (*p*<.05).

Consistent with results from previous studies of video communication, participants preferred the Video condition mostly because of the presence and fidelity of non-verbal cues. For example, "*I could see people's expressions* ($P_{10,4}$)"; "*allows for non-verbal cues + really lets you communicate effectively* ($P_{7,5}$)"; and "*helpful to gauge body language* ($P_{9,2}$)." Video was also considered to be the "*most personal and professional* ($P_{6,4}$)" and the "*most real* ($P_{8,4}$)."

Users liked the Audio condition for its comfort and familiarity. For example, "*Audio is next best + is readily available worldwide* ($P_{7,5}$)"; "*Audio=Not great, but not terrible. Something that we are all used to.* ($P_{6,4}$)"; and "*I'm used to it—it's comfortable* ($P_{7,3}$)." Other benefits included the fact that it was "*easy to use* ($P_{4,2}$)" and "*least distracting* ($P_{7,3}$)." However, they found that in the Audio condition, it was sometimes "*difficult to interject and add thoughts and concerns* ($P_{4,3}$)"; "*was hard to stay focused* ($P_{8,4}$)"; and was "*hard to manage conversation mechanics* ($P_{5,1}$)."

While many of the participants found the Avatar condition to be "*fun*," their main reservations were that it was not professional or serious enough for a business meeting. For example, "*I can't see it being used in a professional setting. I would have a hard time taking a Kinect meeting seriously* ($P_{4,3}$)" and "*Kinect is fun but not for business. The exception would be if you needed to communicate as effectively as possible but had to remain anonymous* ($P_{7,5}$)."

Participants also found the Avatar condition distracting because of "*erratic arm movements and not always picking up who was talking* ($P_{6,5}$)." Users also wanted "*a clearer way to see who is speaking* ($P_{5,4}$)" and more realistic avatars. However, when participants re-evaluated the three conferencing conditions assuming they were "perfect", the ratings for Video (m=9.1) and Audio (m=7.0) did not change significantly (p>.2) whereas the ratings for the Avatar condition (m=6.8) increased significantly (Wilcoxon Z=-4.4, p<.001). As a result, the difference between Avatar and Audio was no longer significant, (p>.9). Besides illustrating the future potential of avatar conferencing as the technology progresses, these responses also suggest a perception that audio and video have "topped out," with little expectation for improved user experiences in video or audio.

4. DISCUSSION AND CONCLUSION

Overall, we found that Video was significantly better in almost all measures of social presence, non-verbal cues and conversational mechanics, and realism than the Audio and Avatar conditions. This result contrasts with Bente et al. [1], who did not find a difference in video compared to other conferencing technologies. We hypothesize that our groups of four who were familiar with each other presented higher interaction demands compared to the dyads of strangers used in their study.

We were surprised that users preferred Audio over the Avatar condition despite ranking Avatar similarly and sometimes higher on the dimensions studied. Survey responses indicated that the workplace context was not a good fit with the cartoon avatars currently offered in Avatar Kinect. This reaction extends prior work [5] which found certain static cartoon avatar icons to be unsuitable for work. It was also interesting to see that users rated the Avatar condition lower on realism (m=4.0) compared to Audio (m=4.7). We surmise that since the participants knew each other, they were able to project a realistic conference experience even in the Audio condition, but that the cartoon avatars actually interfered with their sense of realism. More research is needed to find avatars that are realistic enough to appear professional but avoid negative reactions around the uncanny valley.

Furthermore, some users found the Avatar system distracting because of flaws in the motion tracking. Nevertheless, users did recognize the benefit of a common setting and opportunity for anonymity in avatar conferencing. Also, their ratings for a perfected system show that avatar conferencing has future potential. As this technology matures with more accurate tracking and more professional avatar representations, there is an opportunity for more widespread use of it in workplace settings.

Although we focused on participant perceptions in audio, video, and avatar conferencing, it would also be useful to evaluate the task outcomes with each technology. A comparison of the brainstorming results among the conditions in a future study could offer further insights into the use of avatar conferencing in stimulating creative work.

5. REFERENCES

[1] Bente, G., Rüggenberg, S., Krämer, N.C, Eschenburg, F. Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations, *Human Communications Research*, 34 (2008), 287-318.

[2] Boyle, M. and Greenberg, S. The Language of Privacy: Learning from Video Media Space Analysis and Design, *ACM TOCHI*, 12(2), June 2005, 328-368.

[3] Erickson, T., Shami, N.S., Kellogg, W.A., Levine, D. Synchronous interaction among hundreds: An evaluation of a conference in an avatar-based virtual environment. *ACM CHI 2011*, 503-512.

[4] Espinosa, J.A., Slaughter, S.A., Kraut, R.E., Herbsleb, J.D. Familiarity, complexity, and team performance in geographically distributed software development. *Organization Science*, 2007, 18(4), 613-630.

[5] Inkpen, K. and Sedlins, M. Me and my avatar: Exploring users' comfort with avatars for workplace communication. *ACM CSCW 2011,* 383-386.

[6] Nardi, B. and Harris, J. Strangers and friends: Collaborative play in World of Warcraft. *ACM CSCW 2006*, 149-158.

[7] Nguyen, David and Canny, John, "Multiview: improving trust in group video conferencing through spatial faithfulness", *ACM CHI 2007*, 1465-1747.

[8] Sallnas, E., Rassmus-grohn, K., and Sjöström, C. Supporting presence in collaborative environments by haptic force feedback. *ACM TOCHI*, 7, 2000, 461-476.

[9] Short, J., Williams, E., and Christie, B. *The social psychology of telecommunications*. London: John Wiley & Sons, 1976.

[10] Tang, J.C. and Isaacs, E.A. Why Do Users Like Video? Studies of Multimedia-Supported Collaboration. *CSCW: An International Journal*, 1(3), 1993, 163-196.

[11] Whittaker, S. and O'Conaill, B. The role of vision in face-to-face and mediated communication. In *Video-mediated Communication*, 1997.

[12] Wobbrock, J.O., Findlater, L., Gergle, D., Higgins, J.J. The aligned rank transform for nonparametric factorial analyses using only anova procedures. *ACM CHI 2011*.