A Communication Task in HMD Virtual Environments: Speaker and Listener Movement Improves Communication

Trevor J. Dodds
Max Planck Institute for Biological Cybernetics
trevor.dodds@tuebingen.mpg.de

Betty J. Mohler
Max Planck Institute for Biological Cybernetics
betty.mohler@tuebingen.mpg.de

Heinrich H. Bülthoff
Max Planck Institute for Biological Cybernetics
Dept. of Brain and Cognitive Engineering, Korea University

Abstract

In this paper we present an experiment which investigates the influence of animated real-time self-avatars in immersive virtual environments on a communication task. Further we investigate the influence of 1st and 3rd person perspectives and the influence of tracked speaker and listener. We find that people perform best in our communication task when both the speaker and the listener have an animated self-avatar and when the speaker is in the 3rd person. The more people move the better they perform in the communication task. These results suggest that when two people in a virtual environment are animated then they do use gestures to communicate.

Keywords: Virtual Reality, Virtual Humans, Telecommunication

1 Introduction

Communication is an essential subtask of any collaborative virtual environment (VE), whatever the application (e.g. gaming, urban planning [1], social systems [2]). The rise of ubiquitous motion tracking technologies (e.g. Microsoft's Project Natal) will make telecommunication itself an affordable application of VE technology. With research claiming the control method for gestures and body language is problamatic [3], we focus on naturalistic interaction using motion tracking. In this paper, 'naturalistic interaction' means body motions were captured and mapped onto a self-avatar in real time.

Daft and Lengel's seminal work on the 'media richness theory' identified media capabilities, one of which being the number of communication channels used, and classified media accordingly [4]. Written documents allowed verbal communication, telephone conversations added tone of voice, and face-to-face communication provided the richest medium, with capabilities such as body language and eye contact.

Where/if other types of technology fit into this theory has been the subject of much debate (e.g. [5]), and some electronic communication has outperformed face-to-face [6]. As such, new models have emerged to understand technology mediated communication [7]. Kock has provided an insight from an evolutionary perspective, and argued non-naturalistic interfaces have a higher cognitive load. [8].

Related work has shown that people have a sense of ownership of their avatars in VE, (e.g. the rubber hand illusion demonstrated in VEs [9], and investigations into third-person out-of-body experiences [10]). Further, corresponding

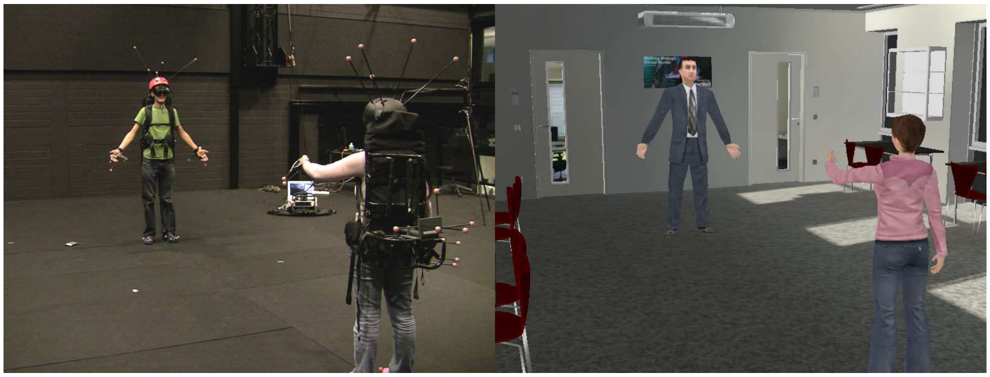

Figure 1: Left: participants wore a total of six tracked objects (2 × hands, 2 × feet, backpack, and helmet). Right: the VE, showing the avatars in the tracked-tracked-3rd person perspective condition.

movements of self avatars have been shown to be more important than the avatar appearance to users' belief that the avatar represents themself [11]. Therefore, our predictions are that people will manipulate their avatars in the environment in a similar way to that of the real world, (i.e. we will see the subconscious gestures that are found to occur alongside speech [12]), and this was tested by comparing the usage of gestures to a real world condition, described below.

2 Experiment

2.1 Method

Participants worked in pairs and played a communication game and had to describe the meaning of words to their partner, who had to guess the word.

The game was played in rounds of three minutes. The study was a 2x2x2 repeated-measures design, and the conditions were changed each round. The conditions varied whether the speaker's avatar was following the movements of the speaker (tracked vs. static), similar for the listener's avatar (tracked vs. static), and whether the camera was in first- or third-person perspective.

In addition, data was collected on a no-vision (black screen) condition and a real world condition (where participants played without a head mounted display, but still wore the markers to collect tracking data). The 10 conditions were presented once for each participant, and were counterbalanced by randomizing the order across pairs.

2.1.1 Participants

8 participants (6 male, 2 female), with a mean age of 26.9 ($SD = 5.3$), took part in the study. Participants were each paired with someone they knew. All participants spoke English as their first language, volunteered for the experiment, gave informed consent, and were paid an honorarium for their participation.

2.1.2 Setup

Participants' body movements were tracked using an optical tracking system (16 Vicon MX13 cameras). Participants each wore six rigid-body objects that were tracked (Figure 1). The objects on their hands were attached across the palm and the wrist. Participants could put the palms of their hands together, but the markers restricted certain gestures close to the body (e.g. participants could not fold their arms).

The virtual reality setup was implemented in Virtools 4.1 from Dassault Systemes. The positions of all joints which were not tracked were calculated using built-in inverse kinematic algorithms, and in addition a calibration was applied which scaled the avatar to the height of the participant. The participants were given a male or female avatar to match their gender.

The environment was an office room (10m length, 6.80m width, 2.77m height) and was symmetrical (left/right walls and front/back walls were the same), apart from the main light source which came from one side only (Figure

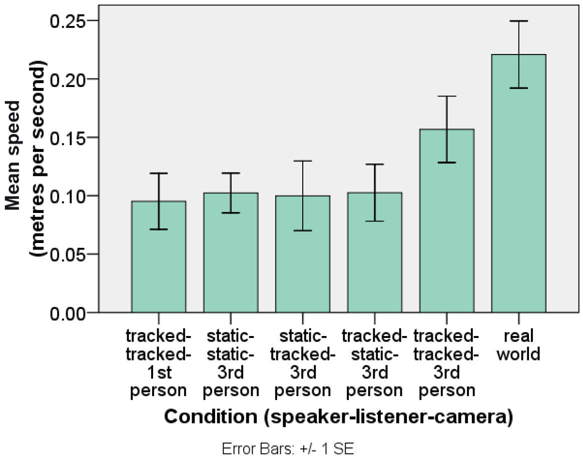

Figure 2: Mean speed of the speaker's dominant hand.

1). Participants stood 4m apart, and each viewed the scene using a light-weight head-mounted display (eMagin Z800 3D Visor) that provided a field of view of 32×24 degrees at a resolution of 800×600 pixels for each eye.

The words to be described in the game were randomly selected from the top 1000 verbs in the British National Corpus.

2.1.3 Procedure

The game was played in rounds of three minutes, with one person as the describer and one person as the guesser in each round. The describer was given words on the screen by the experimenter, and the guesser had to shout out the correct answer — then experimenter provided a new word, and the aim was to get through as many of the words as possible in three minutes.

Describers were not allowed to say what letters were in the word, nor were they allowed to say the word itself, or any derivative, and they were not allowed to use 'rhymes with' clues. When they violated these rules the word was passed and not counted towards success. Similarly, participants were allowed to pass the words, with no consequence except for lost time. Describers were allowed to use gestures, act and mime the word.

Participants were given written and verbal instructions on how to play the communication game, including an example, and played face-to-face before putting on the VE equipment.

2.2 Results

Our two measures per condition were the number of words correct and the amount of movement. For the movement analysis we simply calculated the average speed of the dominant hand per condition. The words correct were automatically stored as a result of the experimenter judgments by button press during the task.

A three-way repeated measures ANOVA was applied to the normalized speed data (log transformed, $D(80) = .06, p = .2, ns$, so results are valid and presented in their non-logarithmic form). There was a significant interaction effect between speaker movement, listener movement and camera perspective, $F(1,7) = 8.57, p = .02$.

Figure 2 shows the mean speed of the dominant hand of the speaker. The conditions represented are first-person perspective with both avatars tracked (left), third-person (middle 4), and real world (right). The third-person perspective saw an increase in the speed, but only when the speaker's avatar *and* the listener's avatar were tracked ($mean = .16, SD = .08$, compared to $mean = .10, SD = .07$ in the first-person tracked-tracked condition). The mean speed was highest in the real world ($mean = .22, SD = .08$).

A bivariate correlation showed a significant positive relationship between the speed of the speaker's dominant hand and the number of words correct, *Kendall's* $\tau = .23, p$ (one-tailed) $= .002$. There was also a significant positive relationship between the speed of the listener's dominant hand and the number of words correct *Kendall's* $\tau = .22, p$ (one-tailed) $= .007$.

Figure 3 shows the mean number of words correct for five of the trials. Participants successfully described a mean of 5.38 ($SD = 2.67$) in the first-person condition with both avatars static, and 10.00 ($SD = 3.55$) in the third-person perspective with both avatars tracked.

3 Discussion

In this experiment we have evaluated how two people perform a communication task, such as conveying the meaning of a word, with their own body gestures mapped in real-time onto a virtual character in an immersive VE. Even with

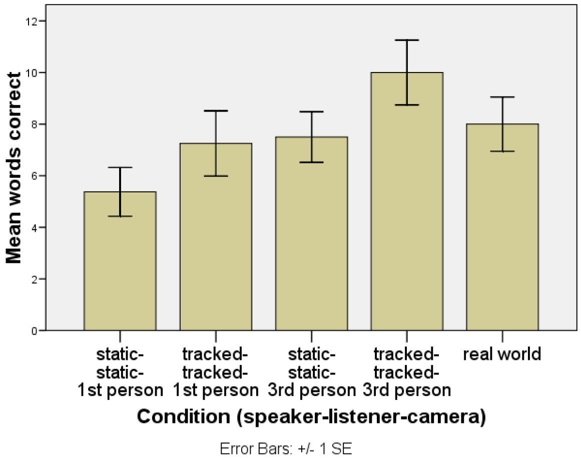

Figure 3: Mean number of words successfully described.

a fully tracked self-avatar people still gesture less in the virtual environment than in the real world. However we find that they gesture most in the VE when the speaker and the listener are both tracked and the speaker is in 3rd person perspective. It is encouraging to see that when people have an articulated self-avatar in a VE they move more, since moving more increases their rate of communication.

Acknowledgements

This work was supported by a post doctoral grant from the Max Planck Society. The authors wish to thank Tobias Meilinger, Stephan Streuber and Stephan de la Rosa for their support.

References

[1] T. J. Dodds and R. A. Ruddle. Using mobile group dynamics and virtual time to improve teamwork in large-scale collaborative virtual environments. *Computers & Graphics*, 33(2):130–138, 2009.

[2] Linden Lab. Second Life. `http://www.secondlife.com/`. (Accessed 26 February 2010).

[3] R. J. Moore, N. Ducheneaut, and E. Nickell. Doing virtually nothing: Awareness and accountability in massively multiplayer online worlds. *Computer Supported Cooperative Work (CSCW)*, 16(3):265–305, 2007.

[4] R. L. Daft and R. H. Lengel. Organizational information requirements, media richness and structural design. *Management Science*, 32(5):554–571, 1986.

[5] M. El-Shinnawy and M. L. Markus. The poverty of media richness theory: explaining peoples choice of electronic mail vs. voice mail. *International Journal of Human-Computer Studies*, 46(4):443–467, 1997.

[6] N. L. Kerr and R. S. Tindale. Group performance and decision making. *Annual Review of Psychology*, 55:623–655, 2004.

[7] A. R. Dennis, R. M. Fuller, and J. S. Valacich. Media, tasks, and communication processes: A theory of media synchronicity. *MIS Quarterly*, 32(3):575–600, 2008.

[8] N. Kock. The ape that used email: Understanding e-communication behavior through evolution theory. *Communications of the AIS*, 5(3):1–29, 2001.

[9] Mel Slater, Daniel Perez-Marcos, H. Henrik Ehrsson, and Maria V. Sanchez-Vives. Towards a digital body: the virtual arm illusion. *Frontiers in Human Neuroscience*, 2, 2008.

[10] B. Lenggenhager, M. Mouthon, and O. Blanke. Spatial aspects of bodily self-consciousness. *Consciousness & Cognition*, 18:110–117, 2009.

[11] Benjamin Lok, Samir Naik, Mary Whitton, and Jr. Frederick P. Brooks. Effects of handling real objects and self-avatar fidelity on cognitive task performance and sense of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 12(6):615–628, 2003.

[12] D. McNeill. *Gesture & Thought*. The University of Chicago Press, Chicago and London, 2007.