



ICT-287760

Vconnect

Video Communication for Networked Communities

Specific targeted research project

ICT – Networked Media

D.6.3 Results of experiments in year 2

Due date of deliverable: 30 November 2013

Actual submission date: 16 January 2014

Start date of project: 1 December 2011

Duration: 36 months

Lead contractor for this deliverable: Falmouth University

Final version, 15 January 2014

Confidentiality status: "Public"

Abstract

This year has been highly productive for Vconnect's user research. The efficacy of three different lay-outs for multi-party video conferencing was evaluated in a highly social setting. Interaction was best supported using a simple tiled display where participants could see each other equally well at all times. A lay-out similar to Google+ Hangout was found to be better suited to business-like discussions in which participants wait their turn to speak.

The lessons learnt were applied to a field trial using SAPO-Campus, an educational social network environment. Analysis is in process but first results are very positive. The study highlights how Vconnect socialisation technologies are finding their way out of the laboratory into the field. Comparing Orchestration with a Static lay-out for video conferencing between three living rooms using HD TV's resulted in findings no strong differences, highlighting that such communication does not translate directly to individuals at computers in multi-party video conferencing.

Galvanic Skin Response was used to monitor audience response during a live theatre performance with the view to visualise remote audience feedback. The perceived quality of the "virtual" microphone was evaluated and we propose a test-battery to gauge quality of video conferencing experience. An overview of related research activities around distributed performance is listed. We detail the user research for the coming year.

Disclaimer

This document contains material, which is the copyright of certain Vconnect consortium parties, and may not be reproduced or copied without permission. All Vconnect consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the Vconnect consortium as a whole, nor a certain party of the Vconnect consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content.

Impressum

Full project title: Video Communication for Networked Communities

Title of the work package: WP6

Document title D.6.3. Results of experiments in year 2

Editor: Erik Geelhoed & Phil Stenton Falmouth University

Workpackage Leader: Phil Stenton

Project Co-ordinator: Peter Stollenmayer, Eurescom

Technical Project Leader: Marian Ursu, Goldsmiths

This project is co-funded by the European Union through the ICT programme under FP7.

Copyright notice

© 2014 Participants in project Vconnect



Abstract of Executive Summary

This year has been a highly productive year for Vconnect's user research.

A key experiment, as part of the socialization use case, was carried out at BT labs. The efficacy of three different lay-outs for multi-party video conferencing was evaluated in a highly social (rather than work-oriented) setting, featuring fast turn-taking, often with more than one participant speaking at the same time. This type of interaction was best supported using a simple tiled display where participants could see each other equally well at all times. A lay-out similar to Google+ Hangout was found to be better suited to business/work-like discussions in which participants wait their turn to speak. The lessons learnt were applied in a field trial at Portugal Telecom (SAPO-Campus) an educational environment within a strong social setting, aimed to afford both a work focus as well as social interaction. Analysis is still in process but first results are very positive. The study highlights how Vconnect socialisation technologies are finding their way out of the laboratory into the field.

Earlier, we evaluated video conferencing between three living rooms (at Goldsmiths University) bridging the technology development between TA2 (Framework 7) and Vconnect. The findings showed the importance of physical location: the setting of two people in a living room in contact with other living rooms using a HD TV does not translate directly to individuals at computers in multi-party video conferencing. In other words, orchestration has no "one size fits all".

Three studies were carried out under the umbrella of instrumenting connected spaces. Galvanic Skin Response was used to monitor audience response during a live theatre performance with the view to visualize remote audience feedback. A lab experiment evaluated the perceived quality of the "virtual" microphone. Another research effort proposed a test-battery to evaluate the quality of video conferencing experience.

We provide a brief overview of related research and networking activities, including obtaining additional funding streams, around distributed performance which highlights how Vconnect's research is deeply anchored into the world of performance, its target audience.

Lastly we list the plans for this year which include a number of performance related public events and a socialization field trial.

The reports, each in different ways, show that in general this second year of Vconnect has seen a very close alignment of the technical work and user research. In addition some of our work has started to get the attention of the wider research community.

Executive Summary

Because this year has seen a plethora of user research activities and extensive reporting, making this document quite large, we decided to write a longer Executive Summary as well as a one page abstract of the Executive Summary (previous page) to make it easier for the casual reader to access the document.

Two laboratory studies (View Mode experiment and View Modes in the Living Room) are reported in great detail and feature certain sections that could be treated as work of reference (for the purpose of dissemination *within* Vconnect, i.e. to describe how to apply sound psychological experimental design and statistical analysis). If a refereed publication represents the top of an iceberg, these two reports show if not the whole then large sections of the iceberg.

Two reports (using Galvanic Skin Response to measure audience response and a proposed test battery to measure Quality of (video communication) Experience) have been written for publication and as a consequence are concise.

The evaluation of the Virtual microphone takes up an intermediary position as this is in a form closer to an external publication.

The report on the SAPO-CAMPUS experiment is, due to time constraints, still in its initial stages and will be reported extensively later this year. All the same there are some important first results to report, as it represents a continuation of the View Mode experiment and it captures how Vconnect socialization technologies are finding their way out of the laboratory into the field.

Lastly we provide a brief overview of related and relevant research and networking activities around distributed performance which highlights how Vconnect's research is deeply anchored into the world of performance, its target audience.

The reports, each in different ways, show that in general this second year of Vconnect has seen a very close alignment of the technical work and user research. In addition some of our work has started to get the attention of the wider research community.

View Mode Experiment

In deliverable D6.2a, SAPO Campus is described as a safe educational on-line environment that serves both social and (educational) work purposes. In the current experiment we evaluate three user interfaces to support ad hoc group video conferencing in our quest to incorporate elements of informal social media usage with more formal work oriented video-conferencing packages. Findings and recommendations of the study have informed the Vconnect integration into SAPO Campus at Portugal Telecom and in a first iteration have been applied in their first experiment/field trial.

At BT labs we conducted nine experimental sessions where in any one session we asked six participants, in six separate rooms, to conduct three separate video conferencing sessions using three different screen lay-outs.



One lay-out was called the Tiled view mode. Participants could see each other, as well as them-selves, depicted in a grid of two rows of three small windows, tiles. This way of displaying the group was thought to provide a good view of the group at all times, but facial expressions (and thereby Telepresence) were expected to be less clear.

The Full Screen view mode would show the person who was talking only, using Vconnect technology this was automatically displayed based on voice activity. Here all the fine details of facial expression would be clearly visible and had the potential deliver a great sense of (individual) Telepresence.

An intermediate view mode was similar to a (Google+) Hangout style of displaying participants, comprising a main window which displayed the current speaker again using voice activity driven Vconnect technology, as well as a row of small windows, displaying the other five (listening) participants,.

Using a repeated measures experimental design that was balanced for order, for each view mode, participants carried out a simple ten minute task, where they had to discuss items that would constitute an ideal holiday, or house, or job. After each view mode, trial participants filled out a short questionnaire asking about their video conferencing experience and at the end of the experiment there was a group discussion asking the participants about what they liked or disliked about the different view modes and they were invited to come up with suggestions how ad hoc/informal video conferencing could be integrated into their social networking activities. Before the experiment we asked participants about their social networking activities and their experiences with video conferencing software.



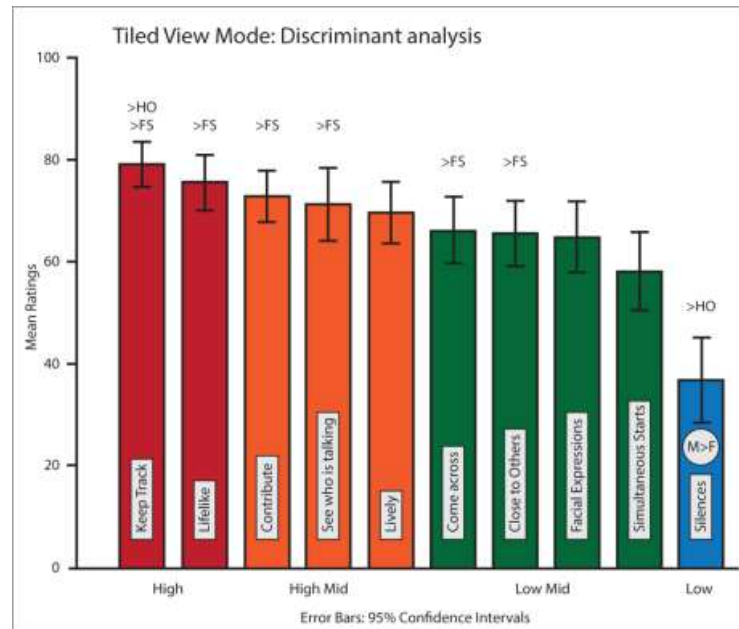
Six participants in soundproof rooms using the Tiled View Mode

We found that most participants were intensive social media users, but the group clearly split into where video conferencing experience was concerned, roughly there was a 60 – 40 split between low (60%) and high (40%) video conferencing software users. The latter group mainly consisted of (slightly) older participants that used video conferencing for work purposes.

All the experimental sessions, irrespective of view mode, were very lively mimicking face to face group interactions, with fast turn taking, participants talking over each other and lots of laughter. In other words the interactions were more social than business like. Having said that, we did find clear differences in how the participants experienced the three view modes.

The Tiled view mode supported this type of lively social interaction well and was generally preferred in particular by the females of the sample. Participants stated that it made everybody equal (the tiles always being the same size), it supported more than one person talking at once and it was possible to monitor the

whole group at all times. It highlighted that monitoring listeners is as important for the conversational flow in social settings as being able to hear speakers.



Mean ratings Tiled view mode

On the Y-axis, in the bar chart above, the mean ratings for the questionnaire results of the Tiled view mode are shown, between 0 (= not at all) and 100 (= very much). The chart also shows which questions were rated significantly higher than the Hangout (HO) and Full Screen (FS) view modes. We identified four separate bands of scores: High (Red), High-Mid (Orange), Low-Mid (Green) and Low (Blue).

Keeping track of the conversation was rated significantly higher in the Tiled view mode than in the other two view modes. Even though the tiles were relatively small in comparison to the size of people who spoke in the Full Screen and Hangout view modes, the participants rated the Tiled view mode significantly more “Lifelike” than in the Full Screen view mode. In the group interviews participants stated that the Tiled view mode felt more “natural”, more like a face to face group discussion, although they would prefer it if the tiles were a bit bigger. They also found it significantly easier to contribute to the conversation; it was easier to see who was talking; participants felt they came across better and felt closer to the other participants than in the Full Screen view mode.

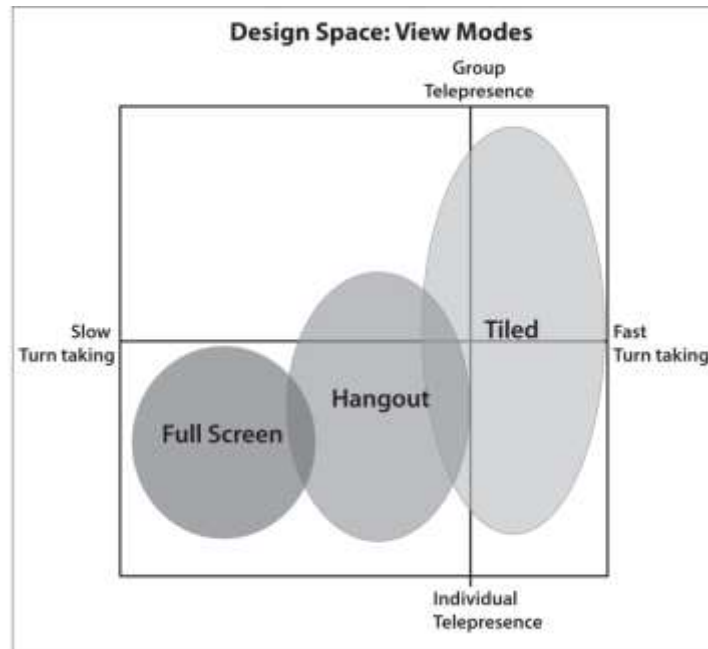
The Full Screen mode was least preferred and all the indications pointed to a group cohesion that was lost which in turn affected participants’ ability to keep track of the conversation.

The Hangout type view mode came a close second (in terms of preference) and was generally preferred by older participants, in particular those who regularly used videoconferencing software for work purposes.

In depth cluster analysis aided in outlining a (two dimensional) design space, see the figure below. The X-axis runs from slow conversational turn taking (as in formal meetings, on-line seminars) to fast and interrupted turn taking, typical of a lively social interaction. The Y-axis ranges from a lay-out more suitable to individual Telepresence (feeling close to one individual) to group Telepresence (feeling close to the group as a whole).

With the integration of Vconnect technology into SAPO-Campus in mind, the Tiled view mode would be highly suitable to support both group and individual telepresence in a fast turn taking scenario, such as one can imagine (in particular female) students get engaged in. The Full Screen view mode would support for

instance an on-line lecture very well. The Hangout interface would be suitable to a more interactive style of educational delivery or for a discussion using digital artefacts, having a mixture of a social interaction and a focus on course work.



Design Space for SAPO-Campus Trials

Based on our findings we recommend a lay-out where one area of the screen would support ad hoc social group interaction and another screen area could be used for sharing educational materials, including on-line seminars. Although Vconnect's user research is more focused on outlining contexts of usage rather than fixing usability issues, as a result of the current study, changes were made very quickly by the technical team to the layout of the Hangout view mode in preparation for the SAPO trials. This responsiveness of the technical team is a good measure for how well user research is integrated into the technology development.

View Modes in the Living Room

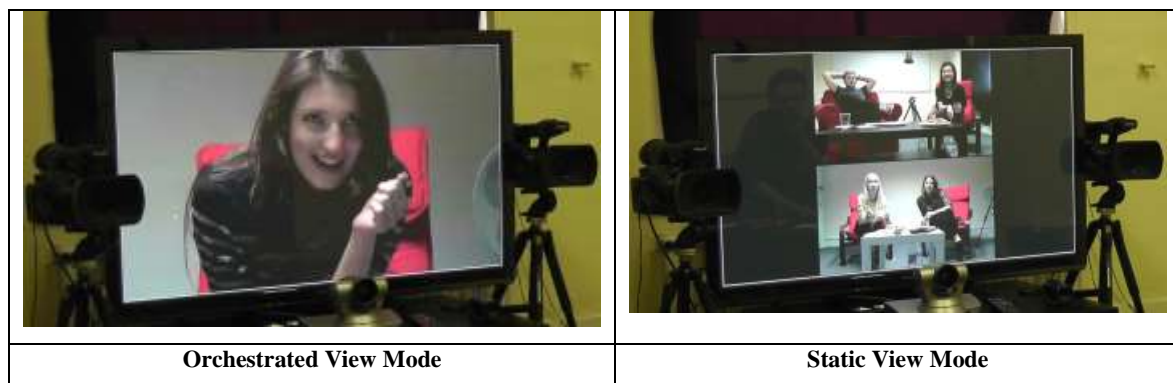
In the second chapter, we report on an experiment we conducted prior to the View Mode experiment above. It functions as a bridge between TA2 research on living room communications (a Framework 7 program preceding Vconnect) and Vconnect's social use case.

At Goldsmiths University, London, three rooms were configured as living rooms with sofas, large TV's and in each room there were three high resolution cameras one of which was a Pan Zoom and Tilt camera (PZT). We conducted four sessions with six participants in each session. In each room there were two participants in video communication with the other two rooms. There were two experimental conditions: One where participants saw a wide shot of the other two rooms (the screen was split) as a static image and the other condition used an automated mix of all three cameras driven by directional voice activity (Orchestration). The PZT cameras were used to deliver close-up shots of a speaker.

We used identical tasks and a similar experimental design as we used for the View Mode experiment, using questionnaires after each experimental trial and group interviews at the end. It is not unreasonable to draw parallels between the Orchestrated and Static view modes of this experiment and the Full Screen and Tiled view modes of the experiment described above.

However, it turned out that the context of a large HD TV-screen and two people sitting next to each other in the same room at a distance of about 2 – 3 meters from the TV-screen resulted in different findings from individuals sitting at a (much smaller) computer screen with headphones on. We did not find significant differences between the two view modes. In the static condition there seemed to be more group awareness facilitating participants to keep track of the conversation. The most striking difference resided in the participants being able to see facial expressions in the Orchestrated condition, enhancing a level of Telepresence of the active speaker.

The two experiments highlight that the physical context of a group’s social communication, i.e. more than one person via a large TV in a living room or one person at a computer with a relatively smaller screen, has a bearing on the efficacy of displaying participants in group communication. In the living room there is something very intimate and evoking empathy about a close-up shot of a speaker or a listener, it somehow compartmentalises a group discussion, giving it more of a feeling of a one to one conversation.



Orchestration using PZT cameras might have the potential to deliver dramatic effect, great Telepresence, in the performance use case. We are currently planning trials for April 2014 at Falmouth University to evaluate the effects, levels of Telepresence using “Vconnected” studios taking advantage of the immersive environments, i.e. the connected CAVE’s, we built. The set-up provides a platform to experiment with various styles of Orchestration and Composition. In addition, as we are instrumenting the performance space with Galvanic Skin Response sensors, it allows us to evaluate audience engagement as well as providing audience feedback. This brings us to the next study.

Sensing the Audience

Returning to the performance use case, we report on a collaboration between CWI and Falmouth University. We used Galvanic Skin Response (GSR) to measure audience engagement. This is part of a series of experiments where we are instrumenting remotely connected performance spaces in such a way that we should be able to visualise remote audience feedback. It is also a nice example of how a sub strand of Vconnect research has moved out of the laboratory into the field and shows great promise for exploitation.

Psycho-physiological measurement has the potential to play an important role in audience research. Currently, such research is still in its infancy and usually involves collecting data in the *laboratory*, where during each experimental session *one* individual watches a video *recording* of a performance. We extend the experimental paradigm by *simultaneously* measuring Galvanic Skin Response (GSR) of a *group* of participants during a *live* performance. GSR data were synchronized with video footage of performers and audience. In conjunction with questionnaire data, this enabled us to identify a strongly correlated main group of participants, describe the nature of their theatre experience and map out a minute-by-minute unfolding of the performance in terms of psycho-physiological engagement. The benefits of our approach are twofold. It provides a robust and accurate mechanism for assessing a performance. Moreover, our infrastructure can



enable, in the future, real-time feedback from remote audiences to online performances. We are currently scaling up the system allowing for simultaneous GSR measurement of larger audiences (up to 40 people simultaneously).

We intend to use the system in future experiments where we compare audience engagement of a co-present and remote audience.

The research is innovative in a relatively new area and has attracted attention from the Human Computer Interaction community. A first short paper won a “Best Paper” award at the 1st International Workshop on Interactive Content Consumption (WSICC), EuroITV 2013, June 2013, Como, Italy. A more extensive paper has been accepted for the prestigious CHI14 conference in Toronto, Canada. This conference receives a great number of submissions but only around 20% of submissions are accepted.

Thus far, the experiments we report on, describe a process of “getting out of the lab” into the “real world”, where Vconnect technology development gets robust enough to be tested in field trials. This last study is a good example of this process.

Of course not all strands of Vconnect technologies are developed at the same pace. One such example concerns the development of the “Virtual” microphone that is still at the laboratory test stage but is no less exciting in scope and potential applicability. The work on the Virtual Microphones also has a good fit under the umbrella of instrumenting the remote (performance and socialisation) space.

Virtual Microphone

The main aim of the virtual microphone (VM) technique is to synthesize the signal of a non-existing (virtual) microphone placed arbitrarily in an acoustic space, which sounds perceptually similar to the signal that would be recorded using a physical microphone located in the same position. Consequently, the subjective evaluation of audio quality and perceptual similarity of such a virtually generated signal to the signal recorded using an analogous physical microphone is of great interest and importance to prove that the technique is right. When changing the VM position within a room with several speakers, the main challenge is to ensure that the generated spatial image of the sound scene (direction and distances to the sources, and the amount of room reverberation) is perceived as closely as possible to the physical microphone recording and that there is no degradation in the audio quality.

For this purpose, we performed a set of three formal listening experiments organized in two sessions of 30 minute duration each. In the first two experiments, the distance and source angle perception were evaluated, whereas in the third experiment, the audio quality and perceptual similarity of the spatial image were compared with reference recordings. The reference stereo signals were recorded using physical (real) microphones at predefined positions in a room. The signals of two distantly placed microphone arrays were also recorded and they were later processed off-line to generate the virtual stereo microphone signals for the evaluation. Thirty listeners participated in the experiments, including both expert and inexperienced listeners. During the assessment, stereo real and virtual recordings were played back over two loudspeakers in one of dedicated listening rooms at Fraunhofer IIS.

Results of these experiments indicate that the spatial image was well preserved in generated stereo signals, often exhibiting no significant difference in perceived distance and angle between physical and virtual microphone recordings. In addition, the audio quality did not suffer from signal degradation in comparison with the reference recordings. Different VM versions were tested and the best performing variants were identified. For example, VM versions based on multichannel signal extraction were shown to outperform VMs based on single-channel enhancement. Although the listeners in general rated the quality of VM processing as good, the VM version capable of real-time performance that utilizes a simplified localizer led to a significantly worse performance for most of tested audio items.

A Quality of Experience Test-bed for Mediated Group Communication

Still under the heading of “instrumenting the space”, but this time purely for research purposes, to provide a measure of efficacy of video conferencing, we report on a test-bed to define Quality of Video Conferencing Experience. As the view modes experiments show, there are many interacting factors that influence the quality of group video communication experience. Here we report a first attempt to develop a test battery evaluating some of such interacting factors, e.g. the effects of delay.

Video-Mediated group communication is proliferating into everyday use, as commercial products enable people to connect with friends and relatives. Although current solutions are impressive, there is still a long way to go to fully support different types of conversations, e.g. remote boardroom meetings vs. social “banter”.

This report (as does the view mode report) argues that the purpose and the context of the conversation are influential factors that are rarely taken into consideration. The aim should be on the development of underlying mechanisms that can seamlessly palliate the effects of networking variances (e.g., delays) and optimise media and connection for every single participant. In particular, our interest is on how to improve remote multi-party gatherings by dynamically adjusting network and communication parameters, depending on the on-going conversation. If we are to provide a software component that can, in real-time, monitor the Quality of Experience (QoE), we would have to carry out extensive experiments under different varying (but controllable) conditions. Unfortunately, there are no tools available that provide the required fined-grained level of control. This first effort in implementing such a test bed provides an experimenter with possibilities for modifying and monitoring network and media conditions in real-time.

SAPO-CAMPUS Vconnect-Integration Trial



Almost directly following on from the view mode experiment at BT labs, both in time and format, trials were conducted at Portugal Telecom, using the integration of the Home Vclient into the SAPO-Campus network.

Even though the time between reporting the results of the view mode experiment (end of August) and the integration of Vconnect’s home client into SAPO-Campus (September) was short, no more than three weeks, the technical team made a number of changes to the user interface in line with the recommendations coming out of the view mode experiment.

To reiterate: SAPO-Campus is a (safe) social network within an educational environment. As such it is well suited to provide features that have both a social and work focus.

The trials were conducted in November and all the data (similar to the view mode experiments) have been collected comprising of post experiment group interviews, questionnaires, logging data and video footage. Due to the holiday period (December) analysis is in the initial stages.

Having learnt from the view mode experiments, the tasks have both social (fast turn-taking in groups) and work oriented characteristics (slower turn taking between individuals) in that initially a group was asked to decide on a conversational topic (fast turn-taking) and then had to discuss (in a more formal manner) different sides of an argument. The trials also evaluated Tiled vs. Hangout view modes.

Twenty five students mostly in the age range 18-25, 17 males and 8 females, took part. First impressions (mainly based on the post-hoc group interviews) are positive. There is a real desire by the participants to integrate (Vconnect) video conferencing into SAPO-Campus.

A full report will be available towards the end of February.

Related activities

One concern for the reviewers of the first yearly review (January 2013) related to our level of involvement with the world of performers and performance; one of our target groups. Of course it helps that both Falmouth and Goldsmiths Universities have Departments of Performing Arts and that experimental as well as developmental (CAVEs) work has involved Falmouth based performance lecturers and students in on-going informal discussions and student projects. In the GSR experiment the actors were closely involved making the GSR sensors “work”, by devising sections that required more or less active, i.e. physical, involvement by the audience.

In addition we presented our work and networked at a conference on mediated communication and performance art. This resulted in establishing some on-going contacts with performance artists experimenting with this type of technology. Moreover we organised a well-attended conference at Falmouth University in this area where artist and technologists met. This included developing close connections with a special interest group of the UK academic broadband JANET around delivering interactive mediated performance. Recently we were successful in an application to JANET performance research arm.

Vconnect involvement also helped Falmouth University in making a successful funding bid to the National Endowment for Science, Technology and the Arts (NESTA) in collaboration with (amongst others) the Cornwall based but nationally known Miracle Theatre to research how to deliver streamed live performances.



Involvement with Miracle Theatre was instrumental in making the recent demo at ICT2013 in Vilnius a success. We demonstrated a remote performance of Samuel Becket’s *Waiting for Godot*, where one actor

resided in Falmouth and another in Vilnius. Using the Vconnect platform to deliver high resolution and low latency video images, they played live to an audience at ICT2013.

In other words our research is by now deeply integrated into this area where technology and the performing arts come together.

Planned Activities

For Vconnect's last year the planning is clear. On the Performance Use Case side, in chronological order, an experiment is planned comparing "home-viewing" of a theatre performance of an edited (orchestrated) with a static version. On 22nd February we will combine forces with a Yamaha's Disklavier performance in four locations providing live and interactive video connections as well as audience evaluations. In April there will be remote performance trials at Falmouth University and a remote performance of Tempest by Miracle Theatre, the culmination of Vconnect's Performance Use Case, is planned for September. Similarly, for the Socialisation Use Case a field trial is planned in the June-July timeframe, using SAPO-Campus where all the work on the Home Vclient is coming together.

Conclusions

This second year has been highly productive and has seen a high level of integration between the technical team and Vconnect's user research. We conclude that the Vconnect system is now coming out of the lab and being applied in real life settings, both in the social and performance use cases. In addition there has been interest in the research community as well as commercial interest in our work and its potential applications.



List of Authors

Professor S.P Stenton & E.N. Geelhoed, Ian Biscoe Falmouth University

Ian Kegel, Peter Hughes – BT Research and Innovation

Marian Ursu, Michael Frantzis, Manolis Falelakis and Vilmos Zsombori, Goldsmiths' College, University of London

Pablo Cesar, Chen Wang, Marwin Schmidt, Simon Gunkel - CWI - Centrum Wiskunde & Informatica, Amsterdam

Konrad Kowalczyk, Nico Faber - Fraunhofer IIS

Pedro Torres – Portugal Telecom

Internal reviewer:

Dennis Dams, Alcatel-Lucent Bell Labs

Table of contents

| | |
|---|------------|
| Abstract of Executive Summary | 3 |
| Executive Summary..... | 4 |
| List of Authors | 13 |
| Table of contents..... | 14 |
| 1 View Mode Experiment 1 | 17 |
| 1.1 Introduction | 17 |
| 1.2 Method..... | 21 |
| 1.2.1 Participants | 21 |
| 1.2.2 Apparatus, experimental rooms, Vconnect technologies and View Modes | 21 |
| 1.2.3 Procedure..... | 24 |
| 1.2.4 Data collection..... | 26 |
| 1.2.5 A brief overview of some of the statistical analysis used in this experiment..... | 27 |
| 1.3 Results | 31 |
| 1.3.1 Interview results | 31 |
| 1.3.2 Questionnaire Results | 38 |
| 1.4 Discussion..... | 83 |
| 2 View Modes in the Living Room | 89 |
| 2.1 Introduction | 89 |
| 2.2 Method..... | 89 |
| 2.2.1 Participants | 89 |
| 2.2.2 Studios, displays, cameras | 89 |
| 2.2.3 Vconnect System..... | 89 |
| 2.2.4 Experimental conditions, order, procedure..... | 90 |
| 2.2.5 Task | 91 |
| 2.2.6 Measures..... | 91 |
| 2.3 Results | 94 |
| 2.3.1 Interviews | 94 |
| 2.3.2 Automated logging | 96 |
| 2.3.3 Questionnaires | 102 |
| 2.4 Discussion..... | 117 |
| 2.5 Conclusion..... | 120 |
| 2.6 References for both View Mode experiments..... | 120 |
| 3 Sensing the Audience..... | 122 |



| | | |
|----------|---|------------|
| 3.1 | Abstract | 122 |
| 3.2 | Introduction | 122 |
| 3.3 | Related Work..... | 123 |
| 3.4 | Pilot Study | 123 |
| 3.5 | Method..... | 123 |
| 3.6 | Results | 125 |
| 3.6.1 | Audience clustering | 125 |
| 3.6.2 | Unfolding of the performance | 125 |
| 3.6.3 | Pre- and Post- Performance Questionnaires | 126 |
| 3.7 | Discussion..... | 127 |
| 3.8 | Conclusion and Future work | 127 |
| 4 | Virtual Microphone – Listening Experiments | 129 |
| 4.1 | Introduction | 129 |
| 4.2 | Aim..... | 129 |
| 4.3 | Method..... | 130 |
| 4.3.1 | Literature review | 130 |
| 4.3.2 | Procedure..... | 131 |
| 4.3.3 | Experimental Setup | 135 |
| 4.3.4 | Stimulus (Speech Signals)..... | 140 |
| 4.3.5 | Participants | 140 |
| 4.3.6 | Statistical Analysis | 141 |
| 4.4 | Results and Discussion | 143 |
| 4.4.1 | Experiment 1 – Distance Perception..... | 143 |
| 4.4.2 | Experiment 2 – Angle Perception..... | 144 |
| 4.4.3 | Experiment 3 – MUSHRA test..... | 145 |
| 4.5 | Conclusions | 149 |
| 5 | A QoE Testbed for Socially-Aware Video-Mediated Group Communication..... | 151 |
| 5.1 | INTRODUCTION | 151 |
| 5.2 | QoE in Video Mediated Communication | 152 |
| 5.2.1 | Factors of QoE in VMC..... | 152 |
| 5.2.2 | Measuring Methodologies | 153 |
| 5.3 | Video-Mediated Conversation-TestBed | 154 |
| 5.3.1 | Video Client for Multiparty Conferencing | 154 |
| 5.3.2 | Support for QoE Factors..... | 154 |

| | | |
|----------|--|------------|
| 5.3.3 | Assessing Feedback..... | 157 |
| 5.4 | DISCUSSION..... | 158 |
| 5.5 | REFERENCES | 159 |
| 6 | Integration Vconnect into SAPO Campus: first experiment..... | 160 |
| 6.1 | Introduction | 160 |
| 6.2 | Method..... | 160 |
| 6.2.1 | Setup..... | 160 |
| 6.2.2 | Participants | 160 |
| 6.2.3 | Experiment | 160 |
| 6.2.4 | Interviews, Questionnaires, Video footage and Automated Logging..... | 161 |
| 6.3 | Results | 162 |
| 6.4 | Preliminary Conclusions | 163 |
| 7 | Relevant other main activities and connections..... | 164 |
| 7.1 | NESTA grant with Miracle Theatre | 164 |
| 7.2 | Vilnius Demo..... | 165 |
| 7.3 | JANET connection and grant application..... | 165 |
| 7.4 | Fascinate Conference – Falmouth University, August 2013 | 166 |
| 8 | Conclusions | 167 |
| 9 | Planning..... | 168 |
| 9.1 | Performance Use Case:..... | 168 |
| 9.2 | Socialisation Use Case | 170 |

1 View Mode Experiment 1

Comparing three different social media group video conferencing lay-outs

Erik Geelhoed¹, Ian Kegel², Michael Franzis³, Niko Faber⁴ and Peter Hughes²

¹ Falmouth University

² BT Labs

³ Goldsmiths University

⁴ Fraunhofer Institute

1.1 Introduction

One of the aims of the Vconnect FP7 research programme is to explore how high quality voice and video, presence and media sharing can be deployed to provide differentiated service offerings. The research focuses on how the provision of new network capabilities such as ‘communication orchestration’ can enable a higher quality, richer ad-hoc video conferencing group interaction in particular through voice driven automated video-editing algorithms that give prominence to the person who speaks during ad hoc group interaction.

More extensively, “Communication Orchestration” is the name given to an automatic decision making process that aims to be “aware” of the conversational flow during a videoconference and is able to represent this flow by controlling participants’ cameras and mixing their content on to the screen at each location. This value proposition may be compared to having an automated ‘personal TV director’ for each location, which chooses the most appropriate shots and presentation to provide the best experience to all participants. Communication Orchestration was conceived in a previous project, TA2 (Together Anywhere, Together Anytime - www.ta2-project.eu), which experimentally showed it to be effective for a group-to-group scenario (in the living room via large TV screens) i.e. when two or more people were present at two or more locations, each of which had multiple cameras from which to select shots.

The Vconnect collaborative project has been developed to address a new opportunity for video communication. This opportunity is emerging at the intersection of two key trends in the consumer market: the rise of online communities, social networks and the rise of video-based communication.

The state of the art in personal audio-visual communication is advancing and specifications such as HTML5 and WebRTC are being implemented to provide a new, open communication platform directly within the web browser. These advances indicate the relevance of Vconnect’s research and are key targets for the exploitation of the project’s results, once browser implementations are more mature. In the interim, the Vconnect platform has been developed to provide a robust and complete communication test bed supporting the orchestration, optimisation and composition features which the project must evaluate in conjunction with real users. For this reason, the first release of the Vconnect platform includes a technical infrastructure which provides similar capabilities to that of Google+ Hangouts today, but extends them through stable implementations of novel server-side components such as the Orchestrator, Optimiser and Video Router. Beyond current consumer videoconferencing solutions, the integrated VClient application supports flexible screen composition under control of orchestration rules.

A detailed description of the Vconnect platform was provided in the deliverable, “Description of Proof of Concept Components, December 2012” [Kegel et al 2012].

User experience varies with different types of orchestration in a social videoconference setting. In the current experiment we evaluate communication orchestration for single-user clients – i.e. each participant joins the conference from a different location and with only one camera, in line with conventional services such as Skype or Google+ Hangouts. Such services have shown that customers are willing to engage in 1:1 or 1: n (where $n < 10$) audio/video chats. Vconnect’s Socialisation Use Case aims to develop these concepts

further, using orchestration as a key component to provide a more compelling experience. It is important to bear in mind that orchestration puts a visual emphasis on the speaker, a vocal individual, at the potential expense of not being able to continuously see and monitor the larger (ad-hoc) group of non-speakers, listening, individuals.

In the literature on video conferencing the importance of silent backchannels (nodding, facial expressions) and short vocal utterances in response to someone speaking both in face to face situations and in mediated communication has been stressed for decades [e.g. Sellen 1992, 1995, Clarke 1991, 1996, O’Conaill et al, 1993, Wilson & Wilson 2005 and Geelhoed 2009]. In other words, in mediated group interactions, whether formal or ad hoc, the person who speaks is important to display of course but to be able to see, monitor the more silent conversational partners at any particular point in time is of equal importance to a healthy conversational flow.

Traditionally (work related) video conferencing sessions between two or more locations had to be booked, arranged similar to a face to face meetings with a location (video conferencing studios), a date, a time, an agenda, supportive documents etc. Such sessions did not take place in an ad hoc manner. In line with the formal nature of such meetings and their intended efficacy, most of the time participants adhere to explicit conversational turn taking, i.e. whilst someone talks the others do not interrupt.

Wilson and Wilson’s seminal work [2005] describes how, when someone is speaking, a listener will anticipate with minute precision when a turn is ending. This implies an entrainment of timing between participants in the conversation. In order for listeners and speakers to show this kind of precision in mutual timing, it is necessary that there is some form of cyclic patterning in the cognitive processes of the speaker that influences the timing of the cognitive processes of the listener, involving endogenous oscillators in the human brain; to put it more metaphysically, this mechanism allows them to share and occupy the same mental space, literally *being on the same wavelength*.

Taking this entrainment one step further, interruptions are a sign of a healthy involved conversation as, before people finish their sentence (a conversational turn) a listener will anticipate where the conversation “is going” and being “entrained” will interrupt a speaker and contribute in turn.

Observing more informal face to face (group) communication between peers, e.g. a group of female teenagers, it is not unusual for there to be plenty of unfinished conversational turns and plenty of interruptions. It might be a requirement for ad hoc video mediated group conversations to technologically support multiple active speakers and to facilitate a quickly shifting, evolving conversational landscape.

Computer based video conferencing packages, Skype, Google+ Hangouts, Facetime, alert a user that colleagues, friends or family are on line and available for a video chat. Currently there are few reports on how such alerts lead to ad hoc, informal video communications.

An initial literature search produced a plethora of technical research papers and only a few recent (2010 and later) human factors publications. E.g. Schreiber et al [2010] carried out an empirical study using “newly formed groups of (three) experts” in Skype communication to carry out a problem solving task. Ickin et al [2012] study the quality of experience of mobile applications in a four week long study. The findings do not throw much light on how such interactions can be made more ad hoc.

Judge and Neustaedter [2010] probe deeper in an interview study how the availability of inexpensive webcams and free video conferencing software such as Google Talk, Windows Messenger, and Skype are used by people as a part of their domestic communication practices. Their findings illustrate the importance of discerning availability and willingness to video conference prior to calling, the need to share everyday life activities in addition to conversation. This seems to highlight that there a tension between a potential demand for ad hoc *informal* group communication, e.g. to share the trivia of daily life and *formal* barriers to organise available slots.

It seems possible both from a design and technology development angle to improve mechanisms for facilitating the occurrence of ad-hoc video mediated gatherings over and above availability alerts.

Here we report on a laboratory experiment where three group communication lay-outs, View Modes, were evaluated. Six rooms were connected via a webcam and microphone using video-conferencing software on the internet. In each room there was one participant and we asked the six participants to engage in a highly lively and social discussion using the three View Modes.

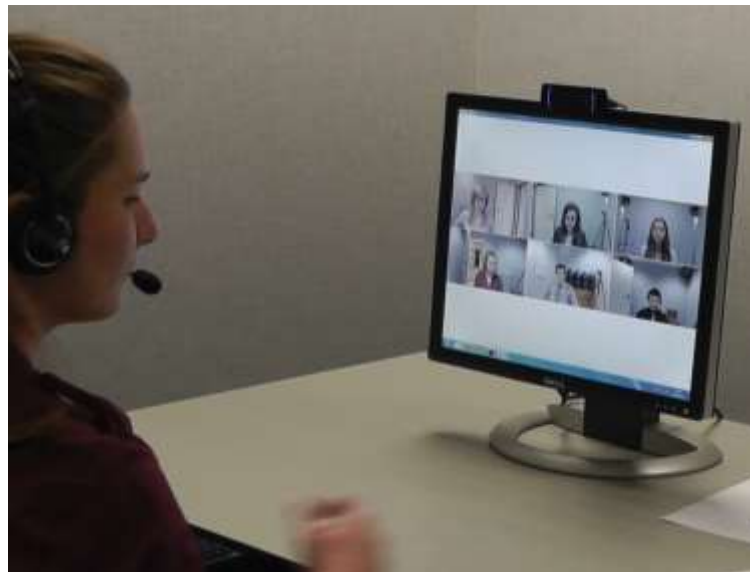


Figure 1.1. Tiled View Mode

In the first condition, the Tiled View Mode (Figure 1.1.) participants could see each other and oneself in a mosaic of tiles, arranged in two rows of three equally sized tiles. This View Mode was thought to afford good group awareness, including being able to identify where the vocal backchannels came from but might be less effective providing feedback through facial expressions. The video composition is static and there is no orchestration, i.e. no matter who is speaking, the monitor always shows the same layout. Therefore, this view-mode can also be considered as the baseline. It allows the user to get a feeling of “who is where” and therefore some form of orientation in the virtual space. A possible benefit of this screen lay-out, view mode, is that participants maintain their position on screen, that is to say, throughout a session they are shown in the same tile, i.e. there is consistency of a participant’s location.



Figure 1.2. Hangout style View Mode

The second condition concerned a screen lay-out similar to the one used in Google+ Hangout (Figure 1.2.), where based on voice detection, the active speaker is displayed in a main window and the five remaining participants are displayed as a row of five tiles. Here facial expressions of the person who talked would be clearly visible with the potential to enhance telepresence of the speaker. In addition group awareness was still accommodated as all participants could still see the rest of the group in the row of smaller tiles. However, there is no consistency in participants' location on screen as they continuously shift position in the row of five small tiles at the bottom of the screen. This might have an adverse effect on group awareness.



Figure 1.3. Full Screen View Mode

In the third condition, again based on voice detection, participants only see the active speaker as a full screen image (Figure 1.3.). This condition was thought to maximise telepresence but group awareness might suffer.

1.2 Method

1.2.1 Participants

A total of 54 volunteers, 18 females (mean age 18.24, SD = 4.07) and 36 males (mean age 20.31, SD = 7.54) took part in the experiment. The graph below (figure 1.2.1.) shows a predominance of participants that were under 20 years of age, as they were recruited from local secondary schools, colleges and universities. All signed a consent form.

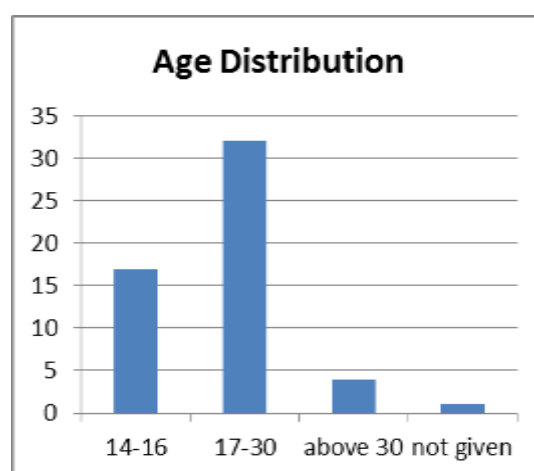


Figure 1.2.1.: Age distribution

1.2.2 Apparatus, experimental rooms, Vconnect technologies and View Modes

The Vconnect platform for the experiment was set up in BT's subjective testing facility in the Orion Building at Adastral Park. The Home VClient software was installed on a Windows PC in each of six silence cabinets, specialist double-skinned rooms which are isolated from external noise. Using these cabinets also prevented acoustic coupling of the rooms, ie. it was not possible to hear other participants other than through the Vconnect platform. The cabinets were connected by a gigabit Local Area Network (LAN), and a connection to the Internet was available for session setup and control. Each silence cabinet was arranged so that the participant sat alone in front of a 19" PC monitor and a webcam. The PC's keyboard and mouse were hidden as the participants were not required to interact with the client itself during the experiment. The setup in each cabinet is illustrated in Figure 1.2.2.1.

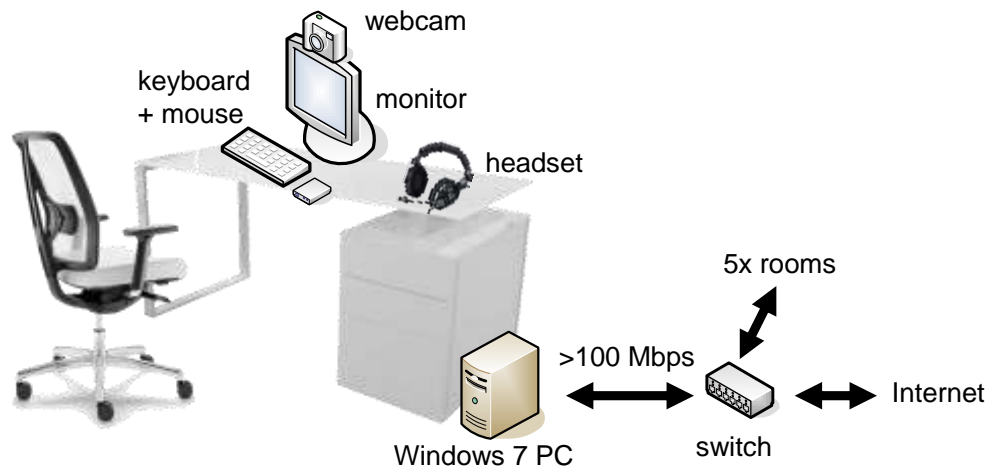


Figure 1.2.2.1: Room setup and components

Windows PC: A PC with an Intel Core-i7 processor running Windows 7 was used to ensure consistent performance.

VClient: The main purpose of the PC was to run the Home VClient, a windows executable, in full-screen mode. The client was configured and started before each experimental run, so the participant did not have to establish a session but simply took their seat in front of an already-running session.

The VClient provided high quality audio and video communication between all participants and rendered the three different view modes on the screen. The view modes could be set remotely using the Orchestrator's web control interface. The VClient also generated Voice Activity Detection (VAD) cues based on the participant's behaviour, performed some basic pre-processing and filtering, and then sent them to the Orchestrator. Figure 1.2.2.2 shows a typical session scenario with 3 participants.



Figure 1.2.2.2: Home VClient on Windows PC

Webcam and Headset: A mixture of Logitech HD Pro C910 or C920 high quality webcams were used. In order to get clean Voice Activity Detection cues the experiment used identical high quality headsets, in this case the Beyerdynamic MMX-2 (figure 1.2.2.3).

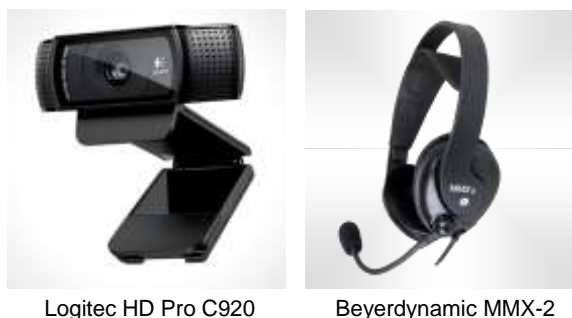


Figure 1.2.2.3: Webcam and headset used

The audio configuration was constant across all view modes and allowed all participants to always hear each other, in line with the standard audio mixing approach adopted in commercial conference bridges. The conference audio was Full-HD quality based on the Audio Communication Engine (ACE) employing the AAC-ELD codec.

Server-side platform configuration

For this experiment, audio and video streams from the VClients were connected to the Video Router and ACE-MCU components, under control of the Optimiser. As explained above, the experiments were conducted under laboratory conditions and as such users did not take part in the process of identifying and establishing a session – hence the Session Manager simply enabled each of the 6 VClients to join the same communication session whenever the experiment system was started up.

During a session, the Orchestrator received Voice Activity Detection cues from each VClient and, depending on the view mode in use, instructed the Optimiser to dynamically select the video streams to be sent to and presented at each VClient. The rules used by the Orchestrator were refined for the purposes of the experiment, and this is described in more detail below.

Orchestrator

The Orchestration Engine essentially consists of two main reasoning components; the Semantic Lifter and the Director. The former component processes and interprets the cues captured by the Analysis components and effectively “lifts” them to infer higher-level interaction events. These are, in turn, processed by the latter in real time to make appropriate editorial decisions and issue corresponding commands to the audio-visual infrastructure. Both the Semantic Lifter and the Director are equipped with prior knowledge, expressed in the form of rules, upon which their decisions are based. These rule-sets are not fixed however; on the contrary, they are meant to change to cater the needs of each specific application, context and/or user.

Below are listed the rules implemented for this experiment. Rules are presented in human language together with a brief explanation and assume that decisions are taken regarding a specific VClient, informed by the status and events occurring in all locations. While the rules may appear relatively simple, they were informed by extensive laboratory testing, and it is important to evaluate their effectiveness alone before further complexity is added in subsequent experiments.

Semantic Lifting Rule

Rule 1: If a person starts to speak for at least X milliseconds then create a ‘significant audio activity event’ for this person.

This rule creates a new significant activity event stating that a certain person is now active. The prerequisite is that the person produces X milliseconds of audio activity, either constant or in chunks. For this implementation X was chosen to be 300ms.

Directing Rules

Rule 1: If ‘significant audio activity’ detected then cut to Close-Up (CU) front at this person.

This rule will result to a “cut”, meaning a change of the displayed image, to a close up front shot of the person.

Rule 2: If current shot duration is less than Y seconds then do not cut.

This rule is imposing a stylistic restriction, preventing a cut if the current shot has not lasted for at least Y seconds. Y was chosen to be 3 seconds.

RTT

The Return Transmission Time (RTT), the standard measure of roundtrip delay, was found to be constant and of the same value between all six rooms and estimated at 650 milliseconds.

Table 1.2.2.1. View Modes

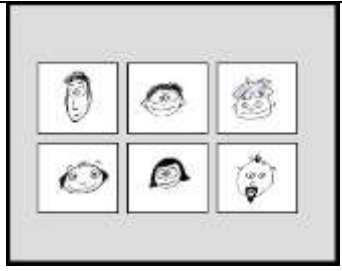
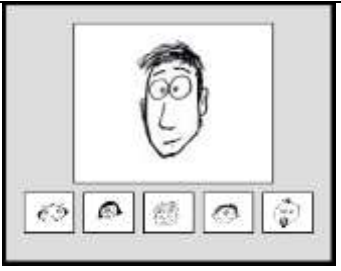
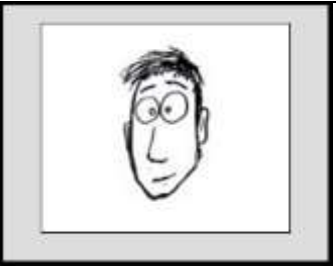
| | | |
|---|---|--|
|  |  |  |
| Figure 1.2.2.4. Tiled | Figure 1.2.2.5. Hangout style | Figure 1.2.2.6. Full Screen |
| Ideal Holiday | Ideal Home | Ideal Job |

Table 1.2.2.1. and figures 2.2.4-6 show the three different lay-out conditions and the topics of the tasks (see below, section 2.3). In the Tiled view mode all participants were displayed in a mosaic of tiles, arranged in two rows of three equally sized tiles. In the Google+ Hangout style view mode, based on voice detection, the active speaker was displayed in a main window and the five remaining participants were displayed as a row of five tiles. In the third condition, again based on voice detection, participants could only see the active speaker as a full screen image. The Hangout style and Full Screen view modes differed in video composition and orchestration, i.e. which participant was seen at which position and at what size, and how this layout was dynamically adapted by the activity of the participants.

1.2.3 Procedure

The experiments were planned to be held over a four day period in late June 2013, and were advertised within BT as well as within a number of local schools and higher/further education establishments via personal contacts. A schedule was finalised according to the availability of the volunteers, resulting in three sessions



being held on each of three days within the planned period. Volunteers (54 in total) came from the following locations:

- Woodbridge School Lower VI Form (private school, East Suffolk) – 18 people
- Hartismere High School Lower VI Form (state school, West Suffolk) – 12 people
- Work experience students aged 14-18 (mixture of schools, East Suffolk) – 9 people
- Cranbrook School Lower VI Form (private school, Kent) – 2 people
- Essex University Masters students, PhD students and staff aged 23-45 – 13 people

Different groups of volunteers were incentivised in different ways depending on the context of their invitation – for example, the Essex University students were interested in a presentation on graduate opportunities within BT and the opportunity to speak to a recent graduate, while the work experience students who responded to an external advertisement were offered a £10 voucher for participating.

An experimental session involved six participants who took part in all three experimental conditions (tiled, hangout, full screen). They were welcomed and briefed about the nature of the experiment and asked to fill out a short questionnaire and consent form. They were then escorted to the subjective testing facility and given a brief tour of a silence cabinet along with appropriate safety instructions. They were also shown how the experiment would be monitored and controlled from a central location, where one of the facilitators would remain in audio (not video) contact throughout the experiment. Each participant was then escorted to an individual cabinet, after which a sound and vision check was carried out to ensure that the communication system was working correctly between all participants and that their headsets were adjusted correctly. Once this was confirmed, the cabinet doors were closed to ensure acoustic isolation.

In the tiled view mode participants had five minutes to generate between themselves at least seven items pertaining to an ideal holiday and after the five minutes had elapsed they were asked as a group to prioritise the items. For the Hangout view mode the topic was “ideal home” and for the full screen view mode this was an “ideal job” (table 1.2.2.1).

The topics for the task are relatively neutral and people of different ages and backgrounds are equally able to carry out the task. In addition there is (practically) no learning curve, as such participants perform equally well in the first as in the third trial. The task is highly repeatable across a number of synchronous and a-synchronous experimental paradigms, i.e. the experimental validity is high. In addition, the tasks elicits highly interactive behaviour around trivial topics in a way that mimics (some) communication through social media and thus we might argue that there is a reasonable ecological validity. It is difficult to strike the right balance between experimental and ecological validity [Clark-Carter, 1997]. Getting the right balance helps to generalise the results to the wider population, i.e. external validity. An added advantage is that the task is relatively short. This helps to keep participants’ experimental fatigue to a minimum.

As task effects have been proved minimal in the past [Geelhoed et al, 2009], it was decided to associate the topics of the task with a specific condition in order to reduce complexity (and possible errors) for the experimental leaders of an already complex experimental configuration.

There were two orders of presentation. In order 1, 30 participants (in five separate sessions of six participants in each) first were exposed to the tiled condition, then to the hangout view mode and lastly to the full screen condition, following a continuously attenuated group representation and enhanced speaker display. In order 2, 24 participants (in four separate sessions of six participants in each) the order was reversed, starting with the full screen view mode, followed by the hangout condition, in turn followed by the tiled condition.

After each trial (View Mode condition) participants filled out a short questionnaire. After all three trials were completed, participants were asked in a three alternative forced choice paradigm (3 AFC) to mark which of the three View Modes they liked best and which one they liked least.

After taking part in the three conditions, participants gathered for a short (20 – 30 minutes) group interview, were debriefed and thanked for their participation.

1.2.4 Data collection

Interviews

After each experimental session the participants were interviewed as a group. The interviews were video recorded and transcribed.

Questionnaires

Before the experimental session, participants were asked to fill out a short questionnaire which asked about their gender, age and how many of the participants in their experimental session they knew. They were asked whether they used the following video conferencing software: Skype, Google Hangout, Apple Facetime or any other package that was not in this list.

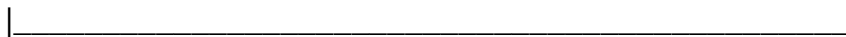
Using graphic rating scales they were asked about the extent of their usage of videoconferencing services and their usage of social networks.

Graphic rating scale questions [Stone et al. 1974] are usually in the form of a continuous line between two extremes (anchors), e.g.

- How often do you use videoconferencing applications, such as Skype?

Not at all

Very



Participants were asked to make a mark between the two extremes (including the two extremes).

The graphic rating scales measured 112 mm. For analysis purposes and to present the results on a scale of 1 – 100, we multiplied the raw data by $100/112 = 0.892857$.

For each View Mode participants were asked an identical set of ten questions about

1. How easy it was to keep track of the conversation
2. How well they felt they came across to the group
3. How well they could see who was talking
4. How lifelike the other participants looked
5. How close they felt to the other participants
6. How well they could see facial expressions
7. How often people started talking at the same time
8. How often there were awkward silences
9. How lively the discussions were
10. How easy it was to contribute to the discussion

The questionnaire data were analysed using SPSS (Statistical Package for the Social Sciences, IBM) providing statistical descriptions; analysis of variance was used to explore differences; correlations and Multi Dimensional Scaling were used to analyse similarities. To compare frequencies, rather than ratings, Chi Squared (χ^2) analysis was used.

1.2.5 A brief overview of some of the statistical analysis used in this experiment

This section provides some background to the statistical analysis approach in the experiments. It is not essential and can be skipped but might be helpful in understanding the questionnaires results.

The only point I would like to make here is that we found a good spread in the ratings for the ten different questions, i.e. participants responded according to their own feelings and opinions and not according to how they might perceive we wished them to respond.

Normal Distribution, Means, Standard deviation, Standard Error of the Mean

In order to help generalising findings from a survey-sample, such as this one, to a larger population, the data are modelled mathematically (statistically) in such a way that we can with some confidence generalise beyond its actual sample size.

As the rating scales lend themselves well to a ratio-scale analysis, the (rather robust) model of the “Normal or Gaussian Distribution” is employed. Named after Carl Friedrich Gauss (1777 – 1855) a typical Gaussian distribution looks like the one depicted below (figure 1.2.5.1.).

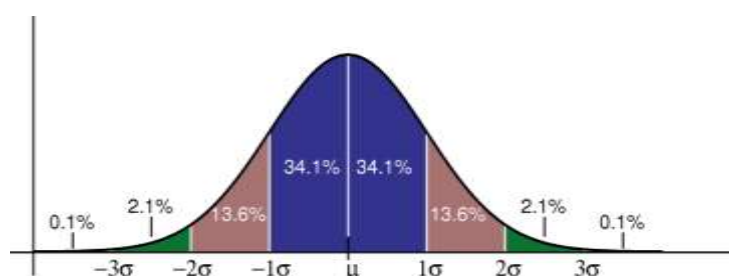


Figure 1.2.5.1: Normal Distribution

An easy way to understand this graph is to imagine the distribution of body-weight of men from low to high, e.g. males between 25 and 35 years in the UK. The mean bodyweight of all the men in that age category in the UK is notified by the Greek letter mu: μ , the population mean.

On both sides of that mean, we see along the X-axis the Greek letter sigma: σ , the Standard Deviation of the population: -1, +1, -2, +2, -3 and +3. Two standard deviations on either side of the mean capture 95.4% of the population, the bodyweight of all men between 25-35 years and only one standard deviation on either side of that mean captures 68.2% (just over two thirds). Thus, one standard deviation on either side of the mean signifies normal body weight. Two standard deviations, a further 27.2%, away from the mean are men who are somewhat “underweight” (on the left) and “overweight” (on the right). The problem cases, where weight is concerned, start to appear at the three Standard Deviation mark, with on the left, the anorexic and on the right the obese.

If we were to carry out a random sample of men’s bodyweight between 25 – 35 years of age, of let’s say, 100 men in that age category and we do this properly along the lines of well tried and tested random sampling techniques, then in all likelihood, we’ll find a similar distribution, with a similar sample mean as the population mean, a similar value for the sample standard deviation as the population standard deviation. Of course this is an idealised example.

In this survey we asked participants to give ratings on a scale of 0 – 100. We can then expect a good 95% of the respondents to fall in within two standard deviations on either side of a mean, and if the mean is close to 50, then we can expect one Standard Deviation to be around 25 (between 20 and 30) as indeed we found for most of the responses to our questions.

A related measure is the Standard Error of the Mean, SEM, signifying confidence margins that a real population mean is within the region around the sample mean for a particular question.

The more normal the distribution is, the more we can generalise to a larger population. The Gaussian model is robust and can withstand some severe violations. One of these violations that can be calculated is called skewness.

Skewness (figure 1.2.5.2): When a normal distribution is leaning towards the left (there are more low values than high values) this results in a positive value for “Skewness” Values of up to ± 2 are acceptable for a distribution to be still considered as normal.

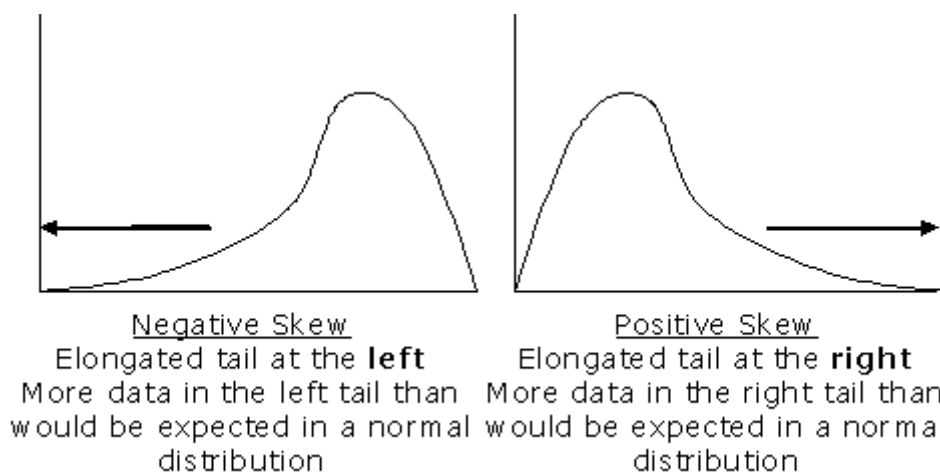


Figure 1.2.5.2 Skewness

Then there is Kurtosis (figure 1.2.5.3). This is measure of the “peakedness” or “flatness” of a distribution. A value close to zero indicates a shape close to normal. A negative kurtosis indicates a flatter distribution and a positive kurtosis a more peaked one. As a guideline values to ± 2 are acceptable for a distribution to be named normal.

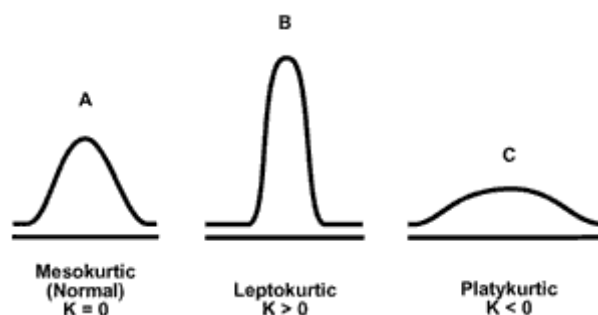


Figure 1.2.5.3: Kurtosis

Levels of significance

When carrying out statistical analysis, the results are accompanied by a level of probability, the so-called p-value. In statistics there are very strict criteria for how to use a p-value and most of the time, we are looking for levels of significance below 5%, i.e. $p = .05$ or less. This indicates the level of confidence we can have that the results did not come about by pure chance alone.

If we find a value of $p = .001$, we can say that if we were to conduct this survey with 54 (different) participants a thousand times then we can expect the results, 999 out 1000, to come out the same. If we were to ignore all statistical research caution, we could say, in, admittedly a rather sweeping, statement, that we can generalise from our sample of 54 to a sample of $(999 * 54 =) 53946$. However ridiculous this sounds, it might be clear that, at least we can generalise considerably beyond the 54 respondents.

In this report occasionally we mention a possible trend when the p - values are below .10 but bigger than .05. The letters n.s. relate to a non-significant finding.

Analysis of Differences

It is desirable to obtain indications whether there are significant differences between the variables with regards to the three experimental conditions. As rating scales are relatively powerful in generating ratio scale data, the tool of choice is the versatile Analysis of Variance (ANOVA). There are a great many flavours of ANOVA and it is particularly powerful in evaluating “within-subject” measures, when for one participant we have more than one score, e.g. participants answering ten questions in a questionnaire, in conjunction with “between subject” measures, e.g. evaluating gender differences and differences in video conferencing experience with regards to how they answered the ten questions.

For nominal and ordinal data here the equally versatile and diverse range of Chi-Squared (χ^2) analyses are used. Chi-Squared analysis will shed light on whether when we observe different frequencies for different variables such differences are to be expected purely by chance alone or whether there is something systematic underlying the observed frequencies. In short Chi-Squared compares what we observe against what we can expect.

In the current study we asked participants to mark which of the three View Modes they liked best and which one least. The ones they liked best received a rank 1 and the one they liked least a rank 3.

Out of a total of 54 participants, 31 participants liked the Tiled view mode best, 22 liked the Hangout one best and 1 liked the Full Screen best. So, our observed frequencies are 31, 22 and 1 for Tiled, Hangout and Full Screen view modes respectively.

Before we started the experiment, we had no way of knowing which would be the preferred view mode (that is why we did the research in the first place). Giving all three view modes an equal chance of $1/3^{\text{rd}}$ (of 54) we could expect 18 subjects voting for each of the view modes. Chi squared analyses then evaluates whether the observed values depart from the expected values in a significant manner. It will not come as a surprise that given that only one participant preferred the Full Screen mode, this departure of the expected frequencies for the three view modes resulted in a highly significant Chi Squared: χ^2 (df 2) = 26.333, $p < .001$.

However, that does not tell us much how the Tiled and Hangout lay-outs compare. In order to do this, we need to carry out two separate Chi-squared analyses.

First we compare the frequency of the rank 1 given to the Tiled view mode, i.e. the Observed frequency is 31, to the combined frequencies of the Hangout and Full screen view modes, i.e. the Observed frequency is 23. The Expected frequencies however are for Tiled $1/3^{\text{rd}}$ of the total frequency (54) = 18 and for non-Tiled $2/3^{\text{rd}}$ of 54 = 36. This analysis produces a highly significant, χ^2 (df 1) = 14.083, $p < .001$, the observed frequency for the Tiled View Mode departs significantly from (higher) what can be expected by chance (of $1/3^{\text{rd}}$) alone.

This exercise was repeated for the Hangout View Mode, the Observed frequency is 22 (Vs. 32 for combined Tiled and Full Screen). This time the expected frequency for the Hangout View Mode is $1/3^{\text{rd}}$ of 54 = 18 resulting in a non-significant χ^2 (df 1) = 1.333.

As such we can with some confidence say that the Tiled view mode is significantly preferred, i.e. departs significantly from what we can expect by chance alone and the frequency of the Hangout style is what we can expect purely by chance.

Analysis of Similarities

Pierson’s Product Moment Correlation is a measure of similarity between two variables, signified by the letter “r”, accompanied by degrees of freedom (df) and a p-value, the level of statistical significance. The degrees of freedom are simply, $N - 1$, i.e. the number of persons in the analysis minus one.

Values for Pierson’s Product Moment Correlation run from $r = +1$, a perfectly positive relationship to $r = -1$, a perfectly negative relationship.

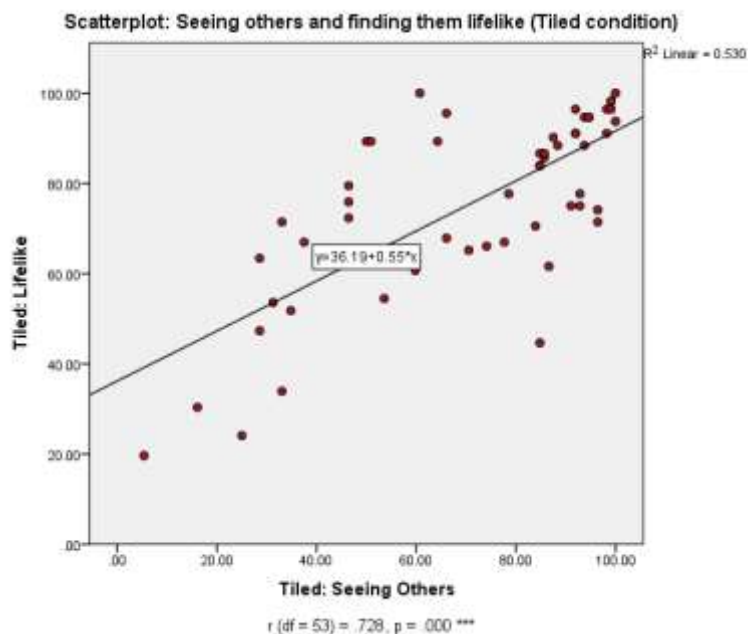


Figure 1.2.5.4.: Scatterplot Tiled Condition: Seeing others and Lifelike

As an example we show in figure 1.2.5.4. along the X-axis, on a scale of 0 (not at all) - 100 (very) how well in the tiled condition participants could see each other and along the Y-axis how lifelike they found each other. There is a strong positive relationship between the two variables ($r (df = 53) = .728, p = .000$). The dots in the scatter-plot signify the data points for each individual for the two variables. The line that runs from the bottom left to the top right, divides the scatter plot in half, i.e. adding up the distances of the dots above the line to that line (at right angles) is equal to the added distances of the dots below the line, and in this fashion does a (relative) good job of signifying the positive relationship between being able to see each other and how lifelike they found each other.

There are some pitfalls using Pierson’s Product Moment Correlation, but these can easily be found in a textbook.

However, Pierson’s Product Moment Correlation is only useful for comparing two variables. In this report, when we wish to depict how a great many variable relate to each other (in a two dimensional plane) we use a cluster analysis technique called Multi Dimensional Scaling.



1.3 Results

1.3.1 Interview results

There were nine experimental sessions followed by group discussions. The distribution of males and females across the sessions is displayed in the table below. The participants are labelled as follows: S refers to sessions, F to females and M to males. For instance S6F3 is the third female participant in session six. The discussion after session 9 was only attended by four males.

Table 1.3.1.: Participants in interview study

| Session | Females | Males |
|---------|-----------|-----------|
| 1 | 4: S1F1-4 | 2: S1M1-2 |
| 2 | 4: S2F1-4 | 2: S2M1-2 |
| 3 | 1: S3F | 5: S3M1-5 |
| 4 | 1: S4F | 5: S4M1-5 |
| 5 | | 6: S5M1-6 |
| 6 | 3: S6F1-3 | 3: S6M1-3 |
| 7 | 3: S7F1-3 | 3: S7M1-3 |
| 8 | 1: S8F | 5: S8M1-5 |
| 9 | | 4: S9M1-4 |

Tiled View Mode

The Tiled View Mode was well received in particular by the female participants. It supported whole group awareness and rapid fire turn taking with more than one person talking.

S1F4: *"I like the first one [Tiled] as well because I just like to be able to see everyone. To see who's talking, because sometimes more than one person would talk at the same time."*

Although for some, it could get a bit too lively.

S3M1: [likes Tiled] *"Where you can see everyone, where you can see how everyone kind of reacts; what someone says and then the bad thing is that, I think people get a bit over excited and like to carry on talking and well, be the main person, sort of thing [very lively] definitely."*

The equal size of the tiles made participants feel that they were equal, included which in turn made it easier to contribute.

S1F2: *"And I think the first one [Tiled] is better because like everyone feels kind of equally included, yeah."*

S3F1: *"I liked the first one [Tiled] because everyone was like the same person."*

S4M1: *"So just the Tiles ones, basically the same cause it was ehm, there was no focus on just the one person, so no one was really being excluded in the conversation, you can see everyone in the same amount of detail."*

It felt more like an actual conversation, mapping on well onto social networking.

S4M2: *“I prefer the Tiles one as well because like when you can see everyone it is more like you can actually talk to them. It is more like an actual conversation.”*

S3M5: *“The first one felt better because it was more like a social, as though you were actually there, and you could see everyone who was talking, as though you were in a circle talking rather.”*

S3M2: *“Yeah I quite liked the first one [Tiled], ehm, I think the first one’s a lot more, better applicable when you’re talking with friends because of lots of people talking at the same time. You can see everybody at once and see people’s reactions to things.”*

It proofed to be important to be able to see oneself.

S1F3: *“Ehm, yeah I like the first one [Tiled] best, because I like the way you can see everybody and you can also see yourself.”*

Paradoxically, in spite of the fast turn taking, the Tiled View Mode, where you could see everyone at all times also caused more awkward silences than the other two View Modes:

S8M2: *“But at the same time with the six tiles, you still got people talking at the same time, there would be an awkward silence, and then two or three people try to talk at the same time.”*

S6F3: *“I thought the last one [Tiled] was best where you could see everyone’s face, so then like we wouldn’t talk at the same time, you were less likely to talk at the same time. So the people, ehm I don’t know, it was just easy to interact as a group discussion.”*

[S8M3]: *“I was also surprised that ehm there’s more awkward silences in, when you could see everyone, rather than when you couldn’t see anyone, whereas the first one [Full Screen]; I think because as soon as there was a tiny bit of silence somebody went in; gotta speak right now. I think that stopped any awkward silences but when you could see everyone, you sort of were looking around waiting for anybody to make a sort of sign for when somebody was going to start and then you know OK no one is going to start, shall I or shall I wait another second and I think that was an issue. [More awkward silences in the Tiles?] Any time you could see everyone there was more awkward silences than with the first one. --- Yeah, I think because you are looking at everyone, you looking for signs that they’re gonna speak and see what everyone’s doing and you sort of wait a little while before just going in there. If you can’t see anyone, as, as soon as there’s a bit of a silence; I was bang in there and I think. But then you get overlap, people trying to talk over each other.”*

Participants expressed that the Tiles should be bigger.

S5M1: *“The first one [Tiled] when there’s ehm all six of us if the screens were bigger so you could see everyone’s faces like a lot clearer and everything that would make it better.”*

S8M3: *“Yeah so the Tiles were too small if they filled up the screen right to each corner then that would be better.”*

Hangout View Mode

The Hangout View mode, in principle at least, also received favourite comments. In particular it afforded more facial details of a speaker whilst still being able to monitor the rest of the group.

S6M3 : *“I thought one with eh as we was talking enlarged and the thumbnails on the bottom was the best [Hangout], because you could easily see who’s talking but you could also see the other people as well. And their reactions.”*



S8M1: *“Yeah the second one [Hangout] because you see the whole group and then who speaks, you can see their facial expression and all of these things.”*

S7F3: *“The best one I thought was the one where you have a big screen [Hangout] and then the people are lined up those who were not talking because then you can focus on someone but you can also still see the other group.”*

The Hangout view mode seemed to help in those situations where you didn't know (many) others.

S6F1: *“I like the Hangout one as well, cause the one with the Tiles sometimes you couldn't see who was talking, 'cause obviously you don't know each other that well, so you don't necessarily know everyone's voices.”*

It seemed to suit business meetings with slower turn taking better.

S1F1: *“I can imagine the second one [Hangout] being quite good if someone is talking for any length of time.”*

S3M2: *“Whereas the second [Hangout] one I think might be more useful like eh business situation like a conference, conference call, because generally people are better mannered and don't interrupt [laughter] ehm so then you get to see the main person and, and be able to see other people's reactions.”*

However the interviews uncovered six problems that would benefit from being addressed:

1. Tiles too small

Although the Hangout view mode also shows other participants, the size of the thumbnails at the bottom of the screen could be increased.

S1M1: *“I like the ehm well the second one [Hangout] with the exception that ehm, maybe a bit of size changing like ehm. I liked being able to see facial expressions better when they were talking. I also liked to be able to see people at the bottom. So, ehm, maybe if the people at the bottom were a bit bigger, and then the person there was still bigger than everyone else, was talking, so you could see them in more detail. Also there were some people that were louder than others, so it would be nice to be able to ehm, individually balance ehm the microphones [group laughter].”*

S2M2: *“Like the second one [Hangout], kind of, it had the extra one and the extra people but they were too small.”*

2. Voice detection Vs. noise detection

Orchestration is driven by voice detection but on occasion Orchestration is determined by noises, e.g. heavy breathing. The following group discussion captures this very well:

S8M2 : *“Maybe that someone was breathing on it.”*

S8M3: *“Maybe your breathing is heavier or something like that, but we were definitely seeing you more than anyone else.”*

S8F: *“Yeah, I was definitely ehm ehm go for the one with the eh eh ribbon down and the who whether who spoke the picture would appear but only if it was ---.”*

S8M1: *“working properly.”*

S1M2: *“--- but I did like the second one [Hangout], because it was good just having the main person up. If, if someone is not doing anything you don't want them, just like taking up half of the screen. which was sometimes the problem with the third one, with someone who had like a slightly background noise or as L. said like breathing and just hog the whole screen for like a couple of seconds and I could see that like being the problem with the second and the third one, if there's, if anyone has any background noise, they'll just be straight up there.”--- “It's just because so many people speak not once, it was just going up down, up down, up down.”*

3. Orchestration lagging

Orchestration was not always able to keep up with the rapid pace of conversational turns.

S8F: *“It didn't work very well --- meaning that when a, when a speaker was speaking it wasn't always ehm reacting. So I kept seeing L all the time, although he didn't speak all the time.--- I would like that [Hangout] but just because it wasn't that reactive. --- If that was improved I would have liked it.”*

S6M1: *“Who was talking had a lag a bit but that was --- as in, who was talking, it was too late coming up who was talking.”*

4. Animation too disruptive

In the Hangout view mode taking a conversational turn was emphasised by an animation where a speaker was relocated from its thumbnail at the bottom of the screen to the main window reserved for a speaker in a way that resembled a genie escaping from a bottle. With the lively conversation this popping up and out became disruptive.

S6F3: *“Like flashing at your face it, like distracting away from the conversation like that.”*

5. Can't see your self

The inability to see your own image made people uncertain whether they would be seen, how they came across.

S6F3: *“You can't see yourself talking [Hangout, Full Screen][group agrees loudly].”*

S6F2: *“Cause when you speak you don't know if everyone's seeing you.”*

6. Location inconsistency

In addition participants in the thumbnails did not keep their position in the tiles at the bottom which was disruptive.

S2M1: *“Yeah I thought you lost track of them when it kept on them [points to big window then to strip of tiles]”* Females: *“Yeah it changed.”* S2F3: *“Couldn't see who was talking.”*

S7F2: *“I was probably, I was like, looking at my notes and looking at the screen, OK, because things keep changing and if I'm focusing on the conversation, if I need to keep focus on what I'm talking, on the subject that we're discussing then it's completely taking you out of focus.”*

Full Screen View Mode



A few male participants liked the Full Screen view mode, mostly because of the clarity of facial expressions.

S3M3: *“That is one of the things of the last one [Full Screen] you got a lot more clarity, the person’s facial expression.”*

S7M1: *“Yeah, for me, I think the best scenario, the best one, ehm, the last [Full Screen] the big picture what that got in all scenarios, yeah the quality of the picture and the voice I find ehm I’m very satisfied with it, very clear.”*

Not having a view on the rest of the group made it very difficult to keep track of the conversation.

S1F2: *“In the last one [Full Screen] where only one person came up, it was kind of hard to, hard to keep in touch with what was going on.”*

In addition the lag of the Orchestration engine was noticed much more in the Full Screen view mode, even though someone put its accuracy at 70%.

S5M3: *“Seventy thirty for showing right, so it was most of the time it was right just a couple of times when it was all with everyone talking. I think sometimes it was a bit late to pick up that that person was talking and sometimes they finished what they were saying before it cut to show them. So the third one [Full Screen] is kind of like you’re kind of singling someone out, it doesn’t feel like a group chatting or because it is just one person on the screen.”*

S1F3: *“Whereas on the last one [Full Screen], it is kind of a bit disconcerting not being able to see the people who were actually speaking and sometimes they would just pop up if someone was just breathing on the ehm microphone. Like, I saw Abi, like, loads of times, she wasn’t saying anything [group laughter].”*

S3M1: *“The one, the last one [Full Screen], where it was like ehm just one person, each time, because like, only kind of telepathic voted who was speaking but it was obviously slightly late, so by the time it will switch, someone else has started speaking. Sometimes it did keep up but then other times, just, slightly slow but it couldn’t be helped.”*

S4M5: *“I don’t know the eh first one [Full Screen] we tried where it was just one big picture you could only, you could hear people but you couldn’t always see them speaking. Yeah because I think it probably picked it whoever was talking and showed a picture of them---- a bit of a delay. Cause someone else might start talking and then it just, it would just keep showing pictures of people when they were not even talking.”*

S4M3: *“And it just showed someone who was like just laughing where someone was actually speaking.”*

It was difficult to monitor the whole group.

S8M2: *“For the first one [Full Screen], what I didn’t like about that because I don’t know, yeah the full screen, I don’t know who or how many people I am talking to; I know it’s six, but I don’t know if maybe someone’s dropped out from their camera, or not actually listen to what I am saying.”*

S7F3: *“OK, ehm, the one I least liked was the last one [Full Screen] because you can’t see everyone else, so you basically can see only the person who’s talking, so you can’t get ehm other peoples interactions from [except from] the guy who talk, so that was my least favourite. Also, there was, a bit of, I don’t know if it’s a technical issue or something, because it is an experiment, with the one where you can see only one person, the*

picture is not always as clear as it should be. Sometimes it gets a bit blurred, you don't spot it too much, but it is enough, so you can notice it. That's basically it."

Out of sight also meant out of mind.

S5M1: *"If one person doesn't talk you kind of forget them [laughter]."*

Participants did not feel part of the group.

S2F4: *"The first one [Full Screen], the first one you couldn't see, so you just thought you were on your own [group in agreement] yeah and then when you could see yourself it was much better."*

S7F3: *"The other thing with the third one [Full Screen] it's the element of, you feel disconnected from the rest of the people, because you know there is other people but you only see one of them, which is the person who is talking, and then if someone else is talking, the element of disconnect-ivity really."*

However, participants could think of scenarios where a Full Screen view mode would be appropriate.

For small numbers of participants:

S3M3: *"I think maybe the third one [Full Screen] would actually be the best if there were three people because it wouldn't have to, you don't have two different images that would flick through, so I think, ehm, it would work really well."*

For mobile phones:

S3M2: *"Maybe on the small screen the third one [Full Screen] would work a bit better I think, because obviously then you end up a bit small."*

When you have multiple screens:

S67M1 : *"Or of you had multiple screens --- If you had multiple screens you could do the first one [Full Screen]. Because then you can have what you see two people."*

Participants' design suggestions

The View Modes experiment seemed to stimulate creative juices, some of them being quite solution oriented.

This one takes advantage of slow turn taking scenarios:

S1M1: [Google hangout] *"For when people are doing like talks or back and forth interviews which you can just in on Google Hangout and you watch it swap between the person interviewing the other person, so it is a lot slower pace than the social sort of back and forth."*

There were two strands as to how to improve the Tiled view mode. The first one was to simplify the interface even further to enhance the group feeling.

S3M2: [Tiled] *"I think maybe if you had like a full screen version of that so the boxes just sort of touched each other ehm so you can see them better I think that might work and also that would make it more maybe ehm more like you were there with them cause if there was, if there weren't like blocks, if there weren't lines between each frame they would all be together, I think that would make it feel like you were just sitting next to each other."*



The second was to highlight the tiles of those who were speaking.

S4M6: *“Or lighting up, lit up around the eh their box when they speak so then you know who’s actually speaking. So you can kind of follow the light pattern and see who’s talking.”*

S5M6 : *”Ehm I think in the first one [Tiled] like if there’s more than a few people talking at once, maybe you’re having like a pop up symbol that shows that that person is the one that is ---.”*

S5M3: [Draws a square in the air] *“Oh yeah like a stock outline.”*

Vconnect intends to facilitate easy joining and leaving of group video chat. The participants had some interesting suggestions.

S3M4: *“If people can come in and out, so like, sometimes, I might be talking to these two guys and she’s busy and then she wants to join the conversation and she just clicks the one, she can join us.”*

There is an awareness that interfaces need to be simple as the discussion after the third session highlights.

S3M2: *“I think like keeping stuff, like, interfaces, keeping them really simple. When things get like overcomplicated where you have like first get that person click on this click on that it just gets too confusing, that will deter people from doing it but if you simplify that you actually like ---.”*

[So you have a circle of friends and you put up a grid]

S3F: *“On my Facebook page, there are green dots, they’re on line.”*

S3M2: *“Like the Facebook thing where it comes up with like messaging there’s a green dot next to my face, that kind of thing.”*

S3F: *“They’re on line, they’re on line.”*

[Then the grid fills up, spontaneously coming in and out]

S3M2: *“I think maybe also if you had it like ehm there’s a method that if you didn’t do it on your phone you would do it on your laptop, you’re talking, say I was talking to D and we would want to talk to Tr, we could click something maybe and send a message to T’s phone to sort of say D and I invited you to voice chat with them sort of thing.”*

Ad Hoc Group Video Chats

Currently our participants are not in the habit of carrying out ad hoc group video chats. There are certain obstacles in organising these.

S1F3: *“But I think if you did have like lots of different people. Like on the first one [Tiled], everyone you talk to, because we often like you’re trying to talk to people on Facebook collectively as a group. And we always trying to organise stuff and then we never ever like get round to it, because some people aren’t on basically when you add them onto it, so you if you have a video call.”*

S1M2: *“To me it’s just eh people, someone’s got dinner, someone’s got some homework to do.”*

S5M3 : *“Well, like it’s hard to get people together really to do that ehm it’s easier just to talk to one person on the internet at once.”*

S6F1: *“I don’t know, you tend to have a Skype if you have like a plan when you want to discuss something or you haven’t seen someone.”*

But when they have a group video chat it is, in the words of one participant, heaps of fun.

S1M1: *“But then the times we have done, ehm, large group videocalls, they have been just heaps of fun. But it’s ehm, it’s, I guess the ability to just drop in and drop out of it. And so you don’t actually have to end the call to leave, you can just say, ehm, everybody else can keep going with it.”*

S8F: *“Family, friends, eh, everything, one to one, and also multiples that was fun.”*

S7F2: *“It was with friends, we were three different stations and eh it went quite good, because usually what you do when you’re Skyping you talk to one station, so it was an exciting to see different stations.”*

They can see the need to get in touch with people you don’t see often, but can’t imagine a need to get in touch again with people you see all day, unless it is about homework (in a SAPO-like scenario) or picking shoes.

S2F3: *“I think maybe if you didn’t see them all day, maybe or they’re in a different country --- faraway.”*

S2F4: *“Unless it’s about homework.”* S2M?: *“If they’re outside the school you didn’t see that often and then you catch up with one person at a time, you could just do it all, now and again.”*

S2F?: *“I don’t know if I’d use it if I saw them every day.”* S2F2: *“What I think like, if you, If I hadn’t seen like, say A. if I hadn’t seen her like all day, or I hadn’t seen her for a week or so, then maybe I’d like to talk to on Skype. So you can see each other and have a laugh about things, ehm, but then if you can see them like every single day more likely to just text someone cause you’ve already seen them, so you don’t get, it sounds weird, so I don’t get annoyed, constantly seeing the same.”* S2F3: *“I find I do most of it in the holidays.”* S2F2: *“Because you’re away from them.”*

S8M3: *“If you could only agree to go into a social media network, somebody would absolutely use that, I think the groups would it. I think if you could somehow ehm, I really like the idea of being able to share browsing as well, I think that’s really intriguing, imagine like the women picking shoes are all picking shoes and stuff --- Have a look at these shoes I’ve seen these, these are better, these are for sale, that sort of thing, I think that would certain take off.”*

1.3.2 Questionnaire Results

First in section 1.3.2.1. we analyse the outcome of the 3AFC (3 Alternative Forced Choice) exercise, i.e. which view mode participants liked best and which one least. At the end of this section there is a short conclusion.

This is followed by the analysis of the pre-experiment questionnaire (section 1.3.2.2.), where we asked about social networking and video conferencing experience and how many of the other participants they knew. This analysis yielded two extra variables, level of video conferencing experience and whether they many (all) or few (none) of the other participants.

We then analyse in section 1.3.2.3. the main questionnaire.

In section 1.3.2.3.1. for each of the ten questionnaire items we analyse whether there were differences between the view modes. Here we take into account gender differences, different levels of video conferencing experience, whether participants knew other participants and differences in the order of presentation of the experimental conditions. We also show how for each question there were significant correlations between view modes.

Given the versatility of ANOVAs, below we include a series of One Way ANOVAs with separately gender, order and video conferencing experience as a factor.

The results of this exploration determines the between subjects factors that are included for fully fledged three Way ANOVAs or sometimes two Way ANOVAs. The decision was partly guided by not having the same number of participants for Order 1 (N = 30) as for Order 2 (N = 24) and we wanted to minimise on

reducing subjects for each cell whilst partitioning the error term. At any rate we used a TYPE IV error term, and as such the results are more conservative.

All of this is designed to untangle findings and explore in more detail from different angles. We use simple versions of Repeated Measures ANOVAs in order to carry out follow up analyses as well as a form of discriminant analysis (section 1.3.2.3.2), where for each view mode we group the variables into separate blocks, e.g. high, mid and low scores.

Finally we analyse similarities using correlations and cluster analysis for each view mode and to some extent between view modes.

1.3.2.1 View Modes Preference

Participants were asked to mark which of the three view modes they liked best and which one they liked least. For all but two participants this resulted in giving a “Rank 1” to the view mode they liked best and a “Rank 3” to the view mode they liked least. The remaining view mode was then given the “Rank 2”.

One participant selected one which was liked best but two which were like least, the latter were both given a tied “Rank 2.5”. Similarly one participant selected one which was liked best but omitted selecting one that was like least, those two unmarked items were also given a tied “Rank 2.5”.

Table 1.3.2.1.1 shows for the 54 participants the total frequencies of the rankings for the three view modes as well as the mean ranks and the graph in figure 1.3.2.1.1. visualises these results.

Table 1.3.2.1.1: Ranks given to View Modes

| Rank | Tiled | | | Hangout | | | Full screen | | |
|------------------|--------------|----|----|--------------|----|----|--------------|----|----|
| | Total | F | M | Total | F | M | Total | F | M |
| 1 | 31 | 14 | 17 | 22 | 4 | 18 | 1 | 0 | 1 |
| 2 | 19 | 4 | 15 | 30 | 14 | 16 | 3 | 0 | 3 |
| 2.5. | 1 | 0 | 1 | 1 | 0 | 1 | 2 | 0 | 2 |
| 3 | 3 | 0 | 3 | 1 | 0 | 1 | 48 | 18 | 30 |
| Mean Rank | 0.778 | | | 0.685 | | | 1.315 | | |

Averaging the ranks it seems that the best results are for the Hangout view mode with the lowest mean rank of 0.685, closely followed by the Tiled view mode and Full Screen trailing behind with a mean rank of 1.315. However, frequencies of rank order data benefit from Chi-squared analyses and this first impression based on the mean rank deserves a closer look.

A rank of 1 (preferred most) was given for the Tiled mode by 31 participants (57.4% of the total number of participants) consisting of 14 females (82% of the 18 females) and 17 males (47% of the 36 males); for the Hangout mode this was 22 times (40.7%) and only once for the Full screen mode.

Carrying out a Chi squared analysis on the frequencies for a rank 1 given, it is no surprise that between the three view modes this results in a highly significant probability, at first glance solely determined by the one participant that ranked Full Screen as 1: χ^2 (df 2) = 26.333, $p < .001$.

It is interesting however, how the Tiled and Hangout lay-outs compare here, i.e. 31 for Tiled Vs. 22 for Hangout. In order to do this, we need to carry out two Chi-squared analyses.

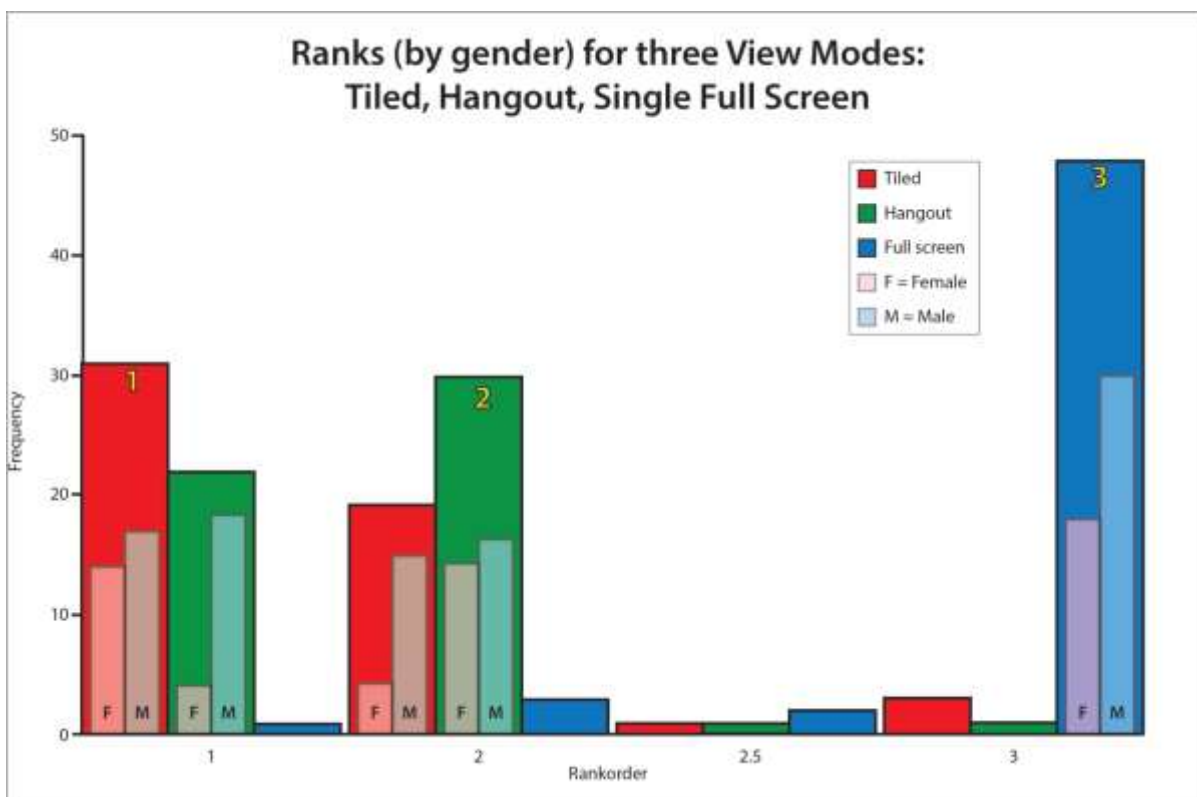
First we compare the frequency of the rank 1 given to the Tiled view mode, i.e. the Observed frequency is 31, to the combined frequencies of the Hangout and Full screen view modes, i.e. the Observed frequency is 23. The Expected frequencies however are for Tiled $1/3^{\text{rd}}$ of the total frequency (54) = 18 and for non-Tiled $2/3^{\text{rd}}$ of 54 = 36. This analysis is highly significant, χ^2 (df 1) = 14.083, $p < .001$, the observed frequency for the Tiled View Mode departs significantly (i.e. is higher) from what can be expected by chance (of $1/3^{\text{rd}}$) alone.

This exercise is repeated for the Hangout View Mode, the Observed frequency is 22 (Vs. 32 for combined Tiled and Full Screen). This time the expected frequency for the Hangout View Mode is $1/3^{\text{rd}}$ of 54 = 18 resulting in a non-significant χ^2 (df 1) = 1.333.

The Hangout mode was ranked second 30 times (55.6%) by 14 females (82%) and 16 males (44.4%), whilst the Tiled lay-out was ranked second 19 (35.2%) times and the full screen view mode three times, resulting in a highly significant Chi squared, again at first glance determined by the three participants that ranked Full Screen second: χ^2 (df 2) = 20.57, $p < .001$.

Repeating the previous two stage Chi-squared analyses we carried out for above for “Rank 1” now for “Rank 2”, there was a non-significant result for the Tiled view mode but a highly significant result for the Hangout view mode χ^2 (df 1) = 13.642, $p < .001$.

The tied ranks are of too low a frequency to consider, but 48 out of the 54 participants (88.9%) ranked the Full screen bottom of the list, consisting of 100% of the females (i.e. 18) and 30 males (83.3%) resulting in a highly significant Chi squared, that is not in need of a follow up analysis: χ^2 (df 2) = 78.63, $p < .001$.


Figure

1.3.2.1.1. View Mode preferences by gender

In figure 1.3.2.1.1., the red bars represent the frequencies of rankings for the Tiled View Mode, the green bars signify the ranks for the Hangout style View Mode and the blue bars the ranks given for the Full Screen View Mode. The frequencies for the Female rankings are shown in transparent pink bars (signified by the letter “F”) and for the males in transparent pale blue bars (signified by the letter “M”). The bars without gender (pink and blue) bars are made up exclusively of male rankings.

To conclude then:

The current analyses indicate that the most preferred View Mode is the Tiled one with 31 participants (out of 54) giving it a rank of “1”; 82% of the females prefer the Tiled view mode. There is a considerable group though (about 40%) that prefers the Hangout style view mode, mostly consisting of males. In spite of average ranks, the Hangout view mode is a close but solid second. It is beyond any doubt that the Full screen view mode is the least preferred one; all the females and over 83% of males prefer the Full screen view mode least.



1.3.2.2 Pre-trial questionnaires

Participants were asked which video-conferencing package they had used. Table 1.3.2.2.1. lists the results. Only one participant had never used any.

The most popular was Skype 50, 93%, had used it. Fewer than half had used Face-Time. There were only 10 users of GoogleHangout, this low number might be partly due (as we learned from the interviews) to age restrictions; a high number of participants were under 18. The category “other” included Oovoo, Tango, Xbox Kinect, Steam Chat and Facebook’s own version of videoconferencing software.

Table 1.3.2.2.1.: Video conferencing software usage

| Software | Frequency |
|-----------|-----------|
| Skype | 50 |
| Face-Time | 26 |
| Hangout | 10 |
| Other | 14 |

In the pre-trial questionnaire participants were asked about the intensity of usage of video-conferencing software and social network, using the following two rating scale questions.

- How often do you use videoconferencing applications, such as Skype?

Not at all Very

- How often do you use social networks, such as Facebook?

Not at all Very

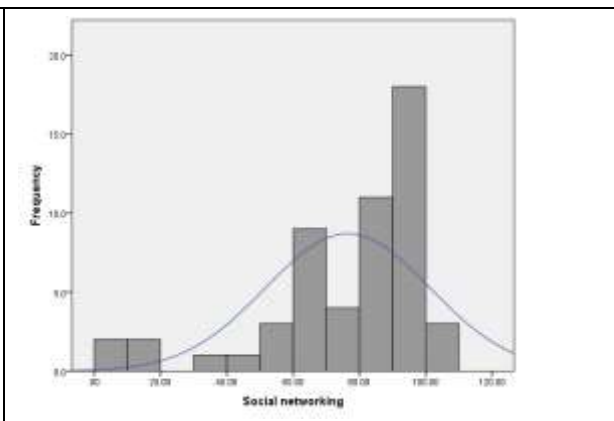
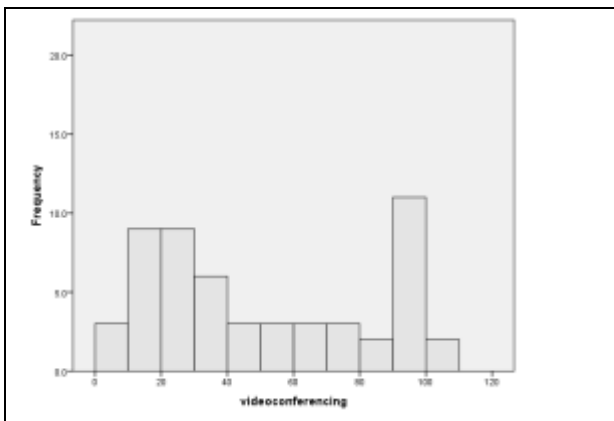


Figure 1.3.2.2.1: Video Conferencing Usage

Figure 1.3.2.2.2: Social Networking

Table 1.3.2.2.2.: Descriptive statistics for videoconferencing and social networking

| | Mean | SD | Median |
|---------------------------|-------|-------|--------|
| Video conferencing | 49.93 | 32.99 | 40.63 |
| Social Networking | 76.65 | 24.41 | |

Comparing the use of video conferencing with the use of social media (ignoring the bimodal distribution of video conferencing) there was a significant effect with the use of social media being significantly higher (repeated measures) ANOVA: $F(1,52) = 22.944, p = .000$.

Based on the median (50% mark) of the video conferencing scores, we performed a median-split, i.e. divided the group into low video conferencing (VC) users (scores of 41 and below) and high video conferencing users. Strictly speaking this was the intermediate score between the 25th and 26th participant. In the tables below in the column VC it will be shown where (for each condition) based on a series of One Way ANOVAs, there were significant differences. The columns VC low (low VC users) and VC high show the means.

1.3.2.3 Main Questionnaire

1.3.2.3.1 Like for Like comparisons

- Question 1: How easy was it to keep track of the discussion?

Not at all Very

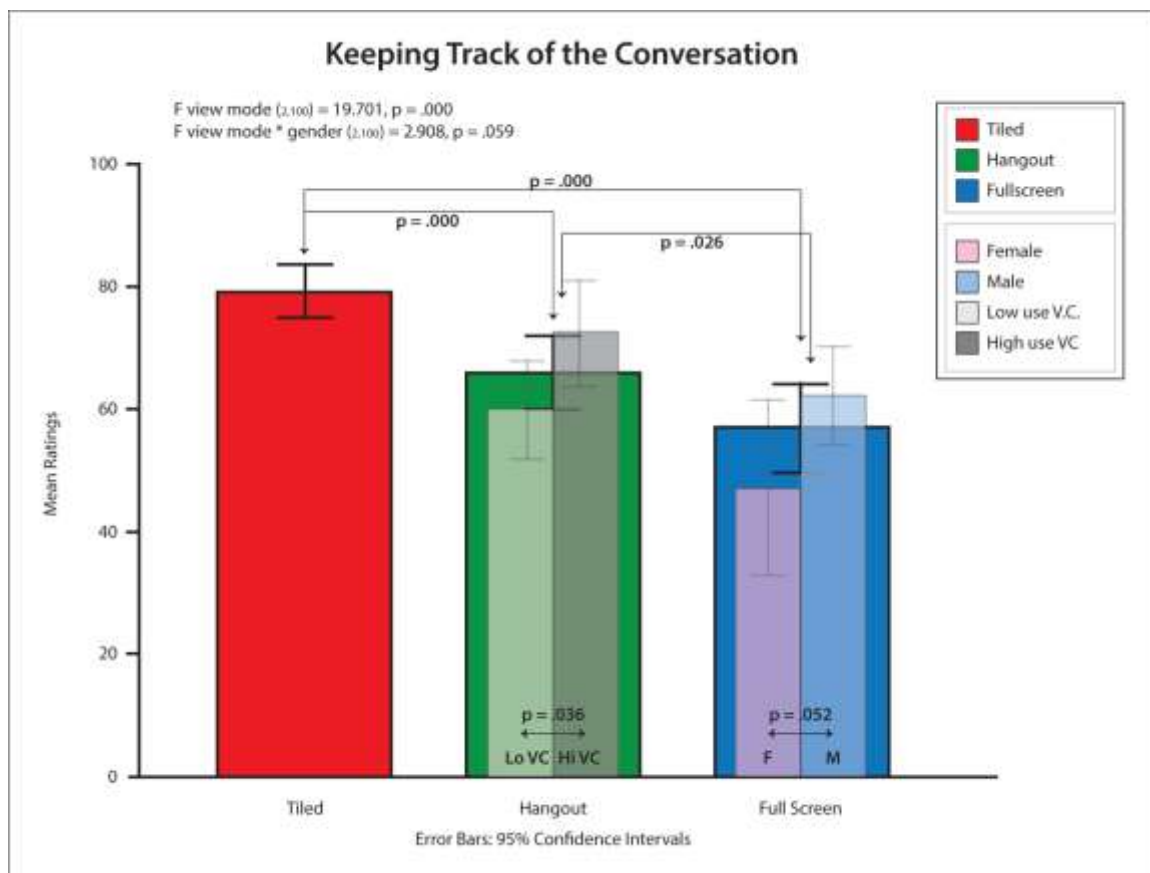


Fig 1.3.2.3.1.1: Keeping Track of the Conversation

Summary:

The ease of keeping track of the conversation (Figure 1.3.2.3.1.1.) is significantly affected by view mode, with the Tiled view mode (red bar) significantly making keeping track easier than the other two view modes. The Full Screen view mode (blue bar) is significantly worse than the other two view modes.

In the Hangout view mode (green bar) participants who are high video conferencing (Hi VC, dark grey bar) users find it significantly easier to keep track than low users (Lo VC, light grey bar) and in the Full Screen view mode males (pale blue bar) find it (almost) significantly easier than females (pink bar).

Correlations between the view modes indicate that those who find it easy to keep track in the Hangout mode also find it easier to keep track in the Full Screen mode.

Detailed Analysis:

Based on exploratory one way ANOVA's a 3-Way ANOVA was carried out where the main factor was the within subjects measure View Mode and two between subjects factors were Gender and Video Conferencing usage (VC) low and high. The latter based on the median split described above in section 1.3.2.2.

Table 1.3.2.3.1.1a. shows the F ratio's, the probability values (p) and for the 3 WAY ANOVA the values of eta squared (η_p^2). Then the results for the one-way ANOVA's used as follow up and in the initial exploration are shown.

Table 1.3.2.3.1.1a. ANOVA: Ease of Keeping Track

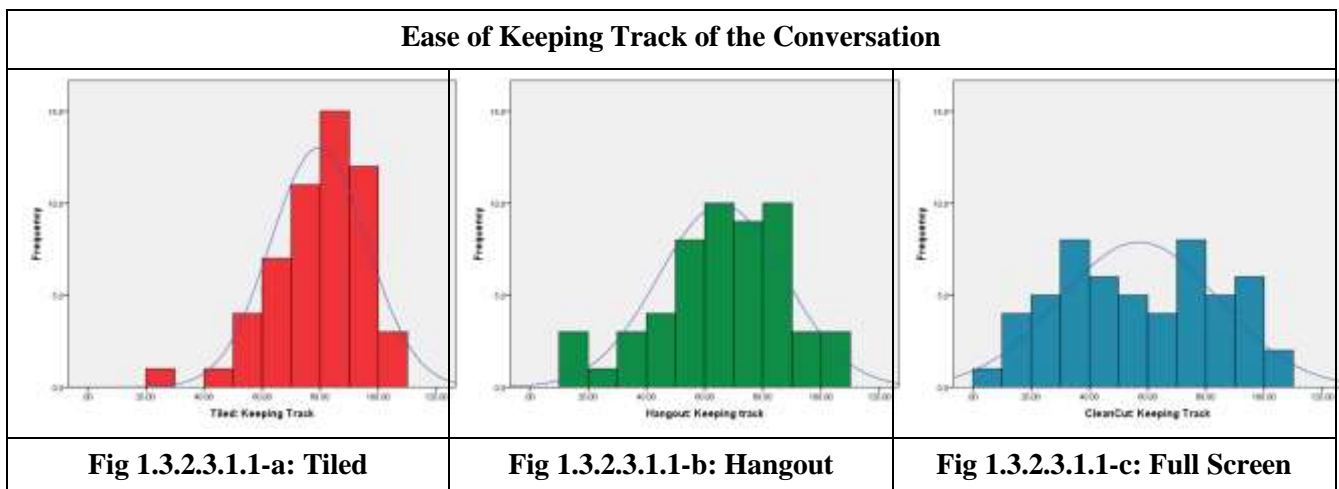
| Factor | F-ratio | p | η_p^2 |
|-------------------------------|------------------------|------|------------|
| View Mode | $F_{(2,100)} = 19.701$ | .000 | .414 |
| View Mode * Gender | $F_{(2,100)} = 2.908$ | .059 | .055 |
| Follow Up (1 WAY) | $F_{(1,53)}$ | | |
| Tiled vs Hangout | 14.341 | .000 | |
| Tiled vs Full Screen | 27.544 | .000 | |
| Hangout vs Full Screen | 5.276 | .026 | |
| Hangout VC | 4.641 | .036 | |
| Full Screen Gender | 3.971 | .052 | |

Table 1.3.2.3.1.1.b. shows the means and standard deviations (SD) for the three view modes. Where there are significant differences within a view mode, i.e. for VC usage and Gender then the means are shown.

Table 1.3.2.3.1.1.b.: Means Keeping Track

| | Mean | SD | F | M | Lo VC | Hi VC |
|--------------------|-------|-------|-------|-------|-------|-------|
| Tiled | 79.28 | 16.21 | | | | |
| Hangout | 66.07 | 22.38 | | | 59.72 | 72.42 |
| Full Screen | 57.06 | 26.62 | 47.12 | 53.86 | | |

Figures 1.3.2.3.1.1-a,b and c show the histograms, the distribution of the scores for the three view modes. There is a peak at the high end of Tiled distribution clearly indicating a high level of concordance (in addition to relative narrow SD) amongst the participants in rating the Tile view mode as easier to keep track. Interestingly the (almost bimodal) histogram for Full Screen view mode shows a small group of participants who find it easy to keep track.





Correlations

Table 1.3.2.3.1.1.d: Correlations between View Modes: Ease of Keeping Track

| | | Tiled | Hangout |
|-------------|-------------|-------|-------------|
| Hangout | r (df = 53) | .147 | 1 |
| | p | n.s. | |
| Full Screen | r (df = 53) | .004 | .318 |
| | p | n.s. | .019 |

Table 1.3.2.3.1.1.c shows the correlations between the View Modes for the ease of keeping track. There is one significant correlation: between the Hangout and Full Screen view modes (figure 1.3.2.3.1.1-d., indicating that those who find it easy to keep track in the Hangout mode also find it easier to keep track in the Full Screen mode.

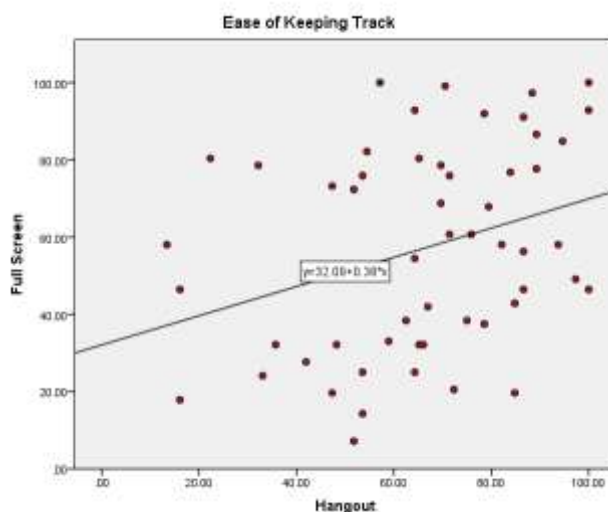
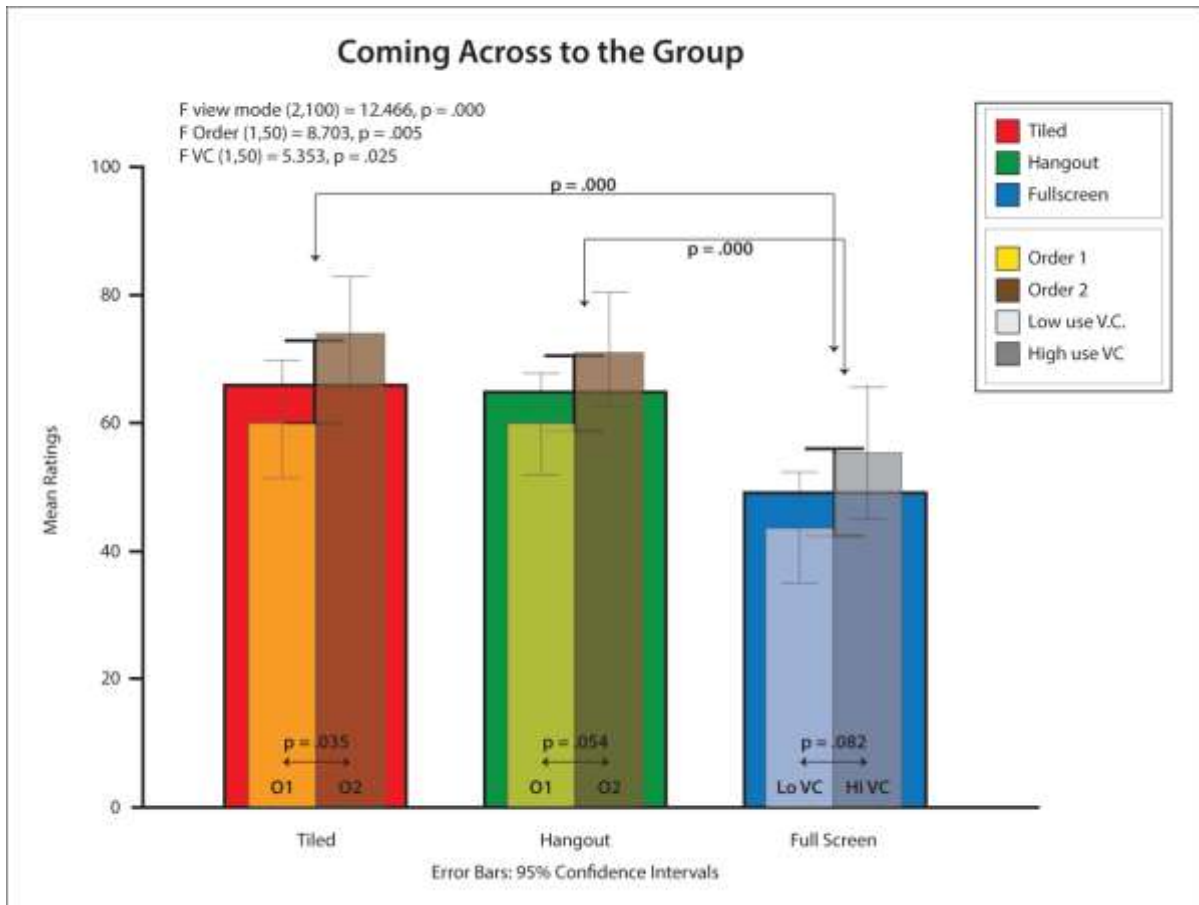


Figure 1.3.2.3.1.1-d: Scatterplot Hangout Vs Full Screen Ease of Keeping Track, r(df 53) = 3.18, p = .019

- Question 2: How well did you feel you came across to the group?

Not at all

Very



Figure

1.3.2.3.1.2: Coming Across to Group

Summary:

In the Full Screen view mode participants feel they come across to others significantly poorer (figure 1.3.2.3.1.2.). For those who are more experienced in video conferencing, this feeling is buffered to some extent; in fact those who feel they come across better in Hangout also do so in the Full Screen view mode. Interestingly, there is a case of negative priming as those who start with the Full Screen (Order 2) seem to feel that they come across better once they have experienced the Hangout view mode and significantly better after they reached the Tiled condition compared to those starting out with the Tiled condition (Order 1). As with the previous question, those who feel they come across better in the Hangout view mode also feel they come across better in the Full Screen mode.

Detailed Analysis:

Based on initial 1-way ANOVAs, we carried out a 3-way ANOVA (Viewmode * Order * VC experience). The significant results and those of the follow up and exploratory one way ANOVAs are shown in table 1.3.2.3.1.2a.

Table 1.3.2.3.1.2a. ANOVA: Coming Across

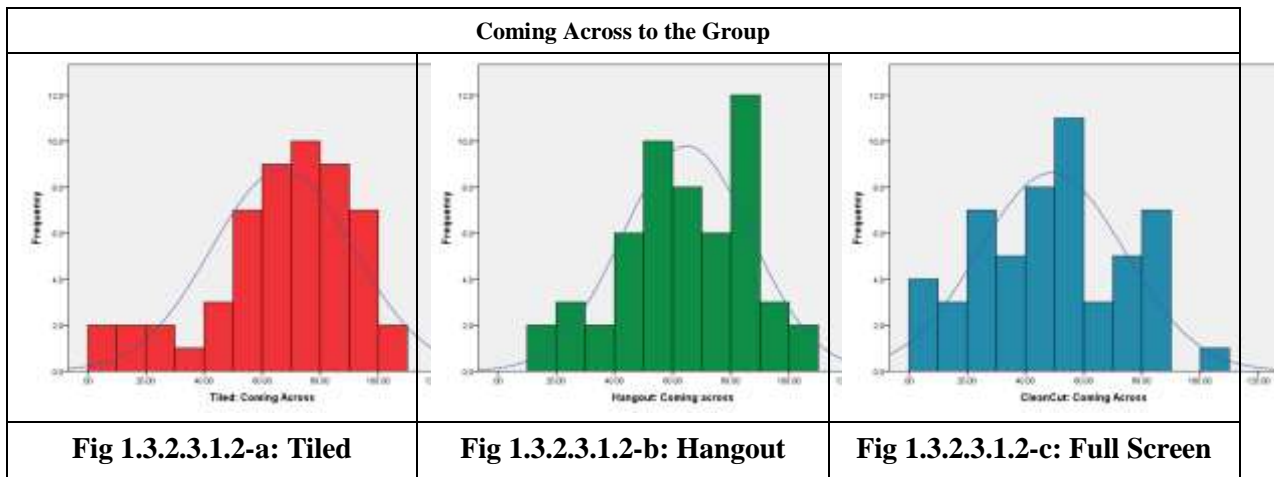
| Factor | F-ratio | p | η_p^2 |
|-------------------------------|------------------------|------|------------|
| View Mode | $F_{(2,100)} = 12.466$ | .000 | .200 |
| Order | $F_{(1,50)} = 8.703$ | .005 | .148 |
| VC | $F_{(1,50)} = 5.353$ | .025 | .097 |
| Follow Up (1 WAY) | $F_{(1,53)}$ | | |
| Tiled vs Hangout | | n.s. | |
| Tiled vs Full Screen | | .000 | |
| Hangout vs Full Screen | | .000 | |
| Tiled Order | 4.641 | .036 | |
| Hangout Order | 3.873 | .054 | |
| Full Screen VC | 3.148 | .082 | |

Table 1.3.2.3.1.2.b. shows the means and standard deviations (SD) for the three view modes. Where there are significant differences within a view mode, i.e. for VC usage and Order then the means are shown.

Table 1.3.2.3.1.2.b.: Means Coming Across

| | Mean | SD | O1 | O2 | Lo VC | Hi VC |
|--------------------|-------|-------|-------|-------|-------|-------|
| Tiled | 66.41 | 24.01 | 60.30 | 74.03 | | |
| Hangout | 64.68 | 21.98 | 59.55 | 61.93 | | |
| Full Screen | 49.14 | 24.72 | | | 43.29 | 54.99 |

Figures 1.3.2.3.1.2-a, b and c show the histograms, the distribution of the scores for the three view modes.



Correlations

Table 1.3.2.3.1.2.c shows the correlations between the View Modes for feeling how participants come across. As with the previous question, there is one significant correlation: between the Hangout and Full Screen view modes (figure 1.3.2.3.1.2-d), indicating that those who feel they come across better in the Hangout view mode also feel they come across better in the Full Screen mode (or vice versa).

Table 1.3.2.3.1.2.d: Correlations between View Modes:Coming Across

| | | Tiled | Hangout |
|-------------|-------------|-------|-------------|
| Hangout | r (df = 53) | .180 | 1 |
| | p | n.s. | |
| Full Screen | r (df = 53) | .241 | .529 |
| | p | .08 | .000 |

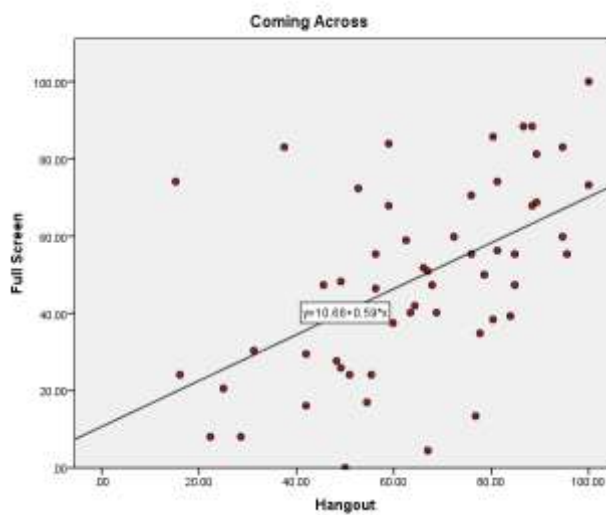


Figure 1.3.2.3.1.1-d: Scatterplot Hangout Vs Full Screen Coming Across, $r(df\ 53) = .529$, $p = .000$

- Question 3: How well could you see who was talking?

Not at all Very

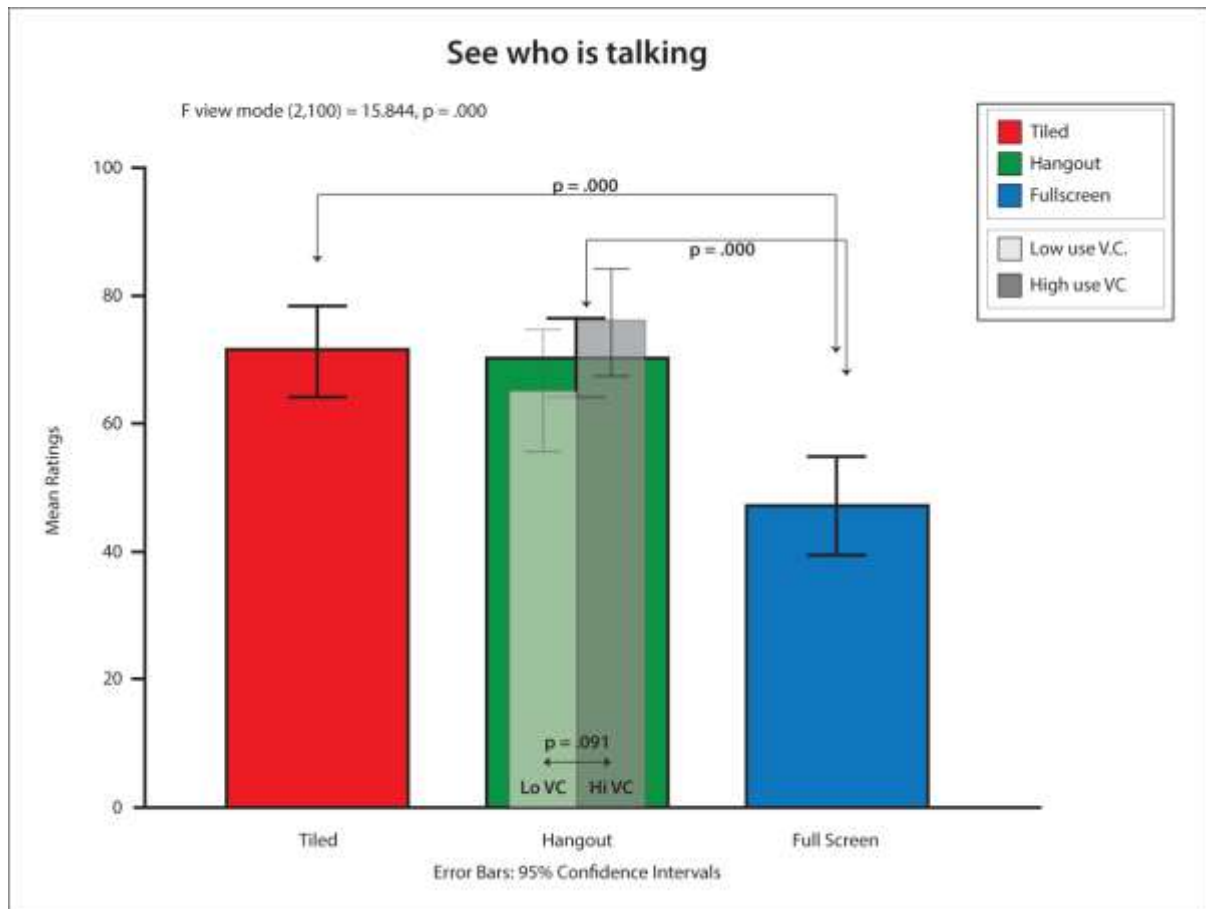


Figure 1.3.2.3.1.3: See who is talking

Summary:

The Full Screen view mode is significantly worse than the other two view modes to be able to follow who is talking. In the Hangout view mode, there is a trend for experienced VC users to be more aware of who is talking. Those who are better able to see who is talking in the Hangout mode also are better in the Full Screen mode.

Detailed Analysis:

Based on exploratory one Way ANOVAs a Two Way ANOVA (View mode * VC experience) was carried out. The significant results and those of the follow up and exploratory one way ANOVAs are shown in table 1.3.2.3.1.3a.

Table 1.3.2.3.1.3a. ANOVA: See who is talking

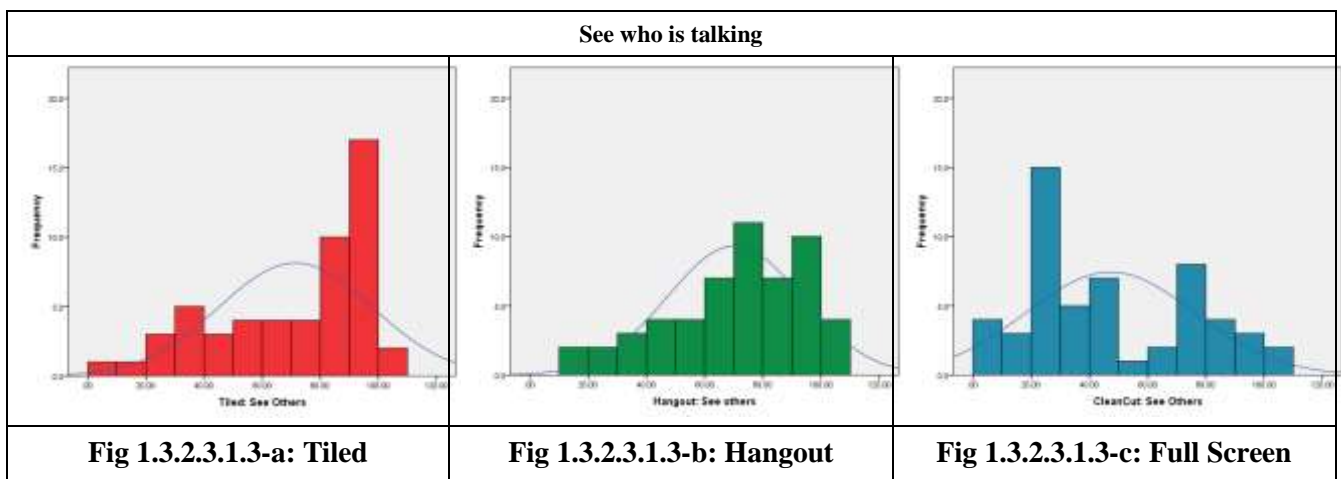
| Factor | F-ratio | p | η_p^2 |
|-------------------------------|------------------------|------|------------|
| View Mode | $F_{(2,104)} = 15.844$ | .000 | .234 |
| Follow Up (1 WAY) | $F_{(1,53)}$ | | |
| Tiled vs Hangout | | n.s. | |
| Tiled vs Full Screen | 19.274 | .000 | |
| Hangout vs Full Screen | 30.135 | .000 | |
| Hangout VC experience | 2.958 | .091 | |

Table 1.3.2.3.1.3.b. shows the means and standard deviations (SD) for the three view modes. For the Hangout mode the means for VC experience are shown.

Table 1.3.2.3.1.3.b.: Means Seeing others

| | Mean | SD | Lo VC | Hi VC |
|--------------------|-------|-------|-------|-------|
| Tiled | 71.44 | 26.47 | | |
| Hangout | 70.33 | 23.16 | 65.01 | 75.66 |
| Full Screen | 47.08 | 28.53 | | |

Figures 1.3.2.3.1.3-a,b and c show the histograms, the distribution of the scores for the three view modes.



Correlations

As with the previous questions the ratings for being able to see the others in the Hangout and Full Screen mode are significantly correlated. In the top right corner in particular is a (smaller) group of participants that gave high ratings in both conditions.

Table 1.3.2.3.1.3.d: Correlations: See Others

| | | Tiled | Hangout |
|-------------|-------------|-------|-------------|
| Hangout | r (df = 53) | -.016 | 1 |
| | p | n.s. | |
| Full Screen | r (df = 53) | -.097 | .289 |
| | p | n.s. | .034 |

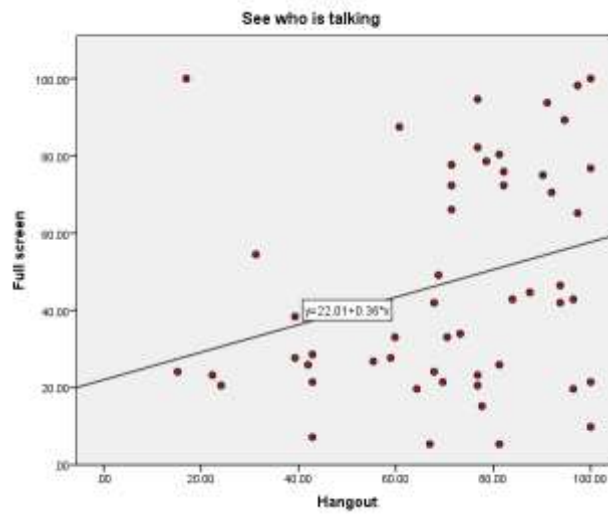


Figure 1.3.2.3.1.3-d: Scatterplot Hangout Vs Full Screen See who is talking, $r(53) = .289$, $p = .034$

- Question 4: How lifelike were the other people?

Not at all

Very

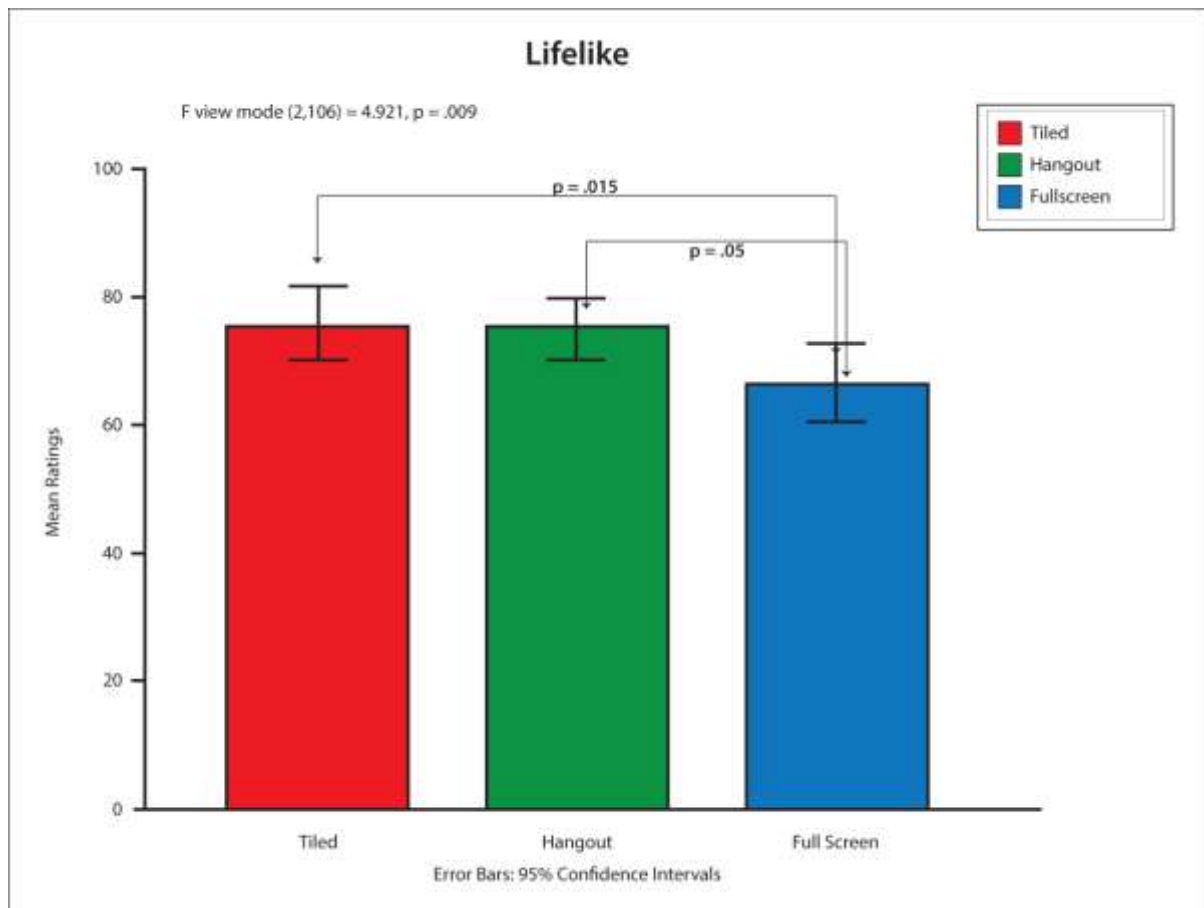


Figure 1.3.2.3.1.4: How Lifelike Participants were

Summary:

Ratings for how lifelike people seemed are higher in the Full screen view mode compared to the previous questions. However, as can be seen from the narrow standard deviations, participants are closely agreed and still perceive that the Hangout and Tiled view modes perform significantly better. this question might have been ambiguous where the Full screen view mode is concerned. Some gave a high rating in the Full Screen view mode possibly having in mind the full screen view of those who talked whereas others might have answered this question because of the lack of view of those that did not talk. Again there is a close relationship between the Hangout and Full screen view modes.

Detailed analysis:

Because in the exploratory one way ANOVAs for gender, order and VC experience, there were no effects for “Lifelike”, a one way Repeated Measures ANOVA (viewmodes) was performed. Details are in tables below.



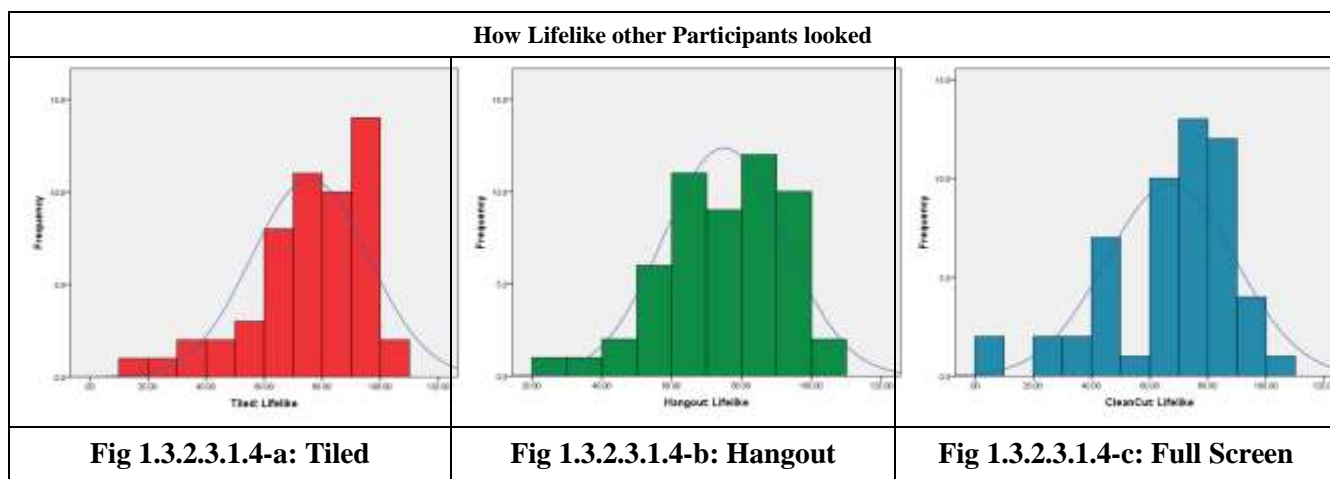
Table 1.3.2.3.1.4a. ANOVA: Lifelike

| Factor | F-ratio | p |
|------------------------|-----------------------|------|
| View Mode | $F_{(2,106)} = 4.921$ | .009 |
| Follow Up (1 WAY) | $F_{(1,53)}$ | |
| Tiled vs Hangout | | n.s. |
| Tiled vs Full Screen | 6.267 | .015 |
| Hangout vs Full Screen | 8.637 | .005 |

Table 1.3.2.3.1.3b. Means: Lifelike

| | Mean | SD |
|-------------|-------|-------|
| Tiled | 75.79 | 20.15 |
| Hangout | 74.98 | 17.42 |
| Full Screen | 66.61 | 21.86 |

From the histograms below it seems that this question might have been ambiguous where the Full screen view mode is concerned. Some gave this a high rating (see peak towards the high end) possibly having in mind the full screen view of those who talked whereas others might have answered this question with the lack of view of those that did not talk in mind.



Correlations

Again there was a highly significant correlation between the Hangout and Full Screen view modes.

Table 1.3.2.3.1.4.d: Correlations: Lifelike

| | | Tiled | Hangout |
|-------------|-------------|-------|-------------|
| Hangout | r (df = 53) | .247 | 1 |
| | p | .071 | |
| Full Screen | r (df = 53) | .180 | .452 |
| | p | n.s. | .001 |

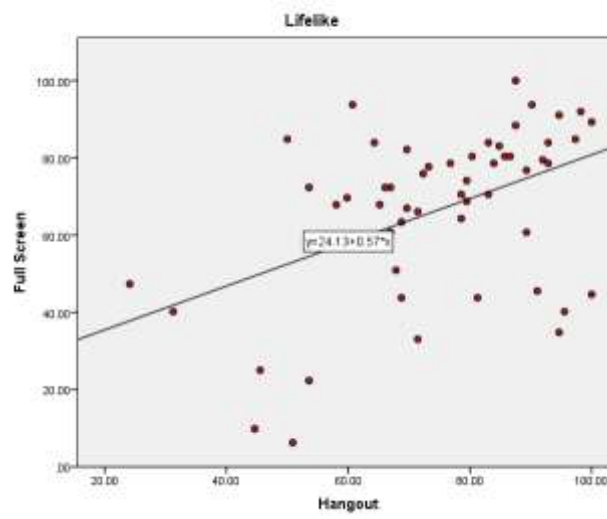


Figure 1.3.2.3.1.3-d: Scatterplot Hangout Vs Full Screen Lifelike, $r(53) = .452$, $p = .001$

- Question 5: How close did you feel to the other people?

Not at all

Very

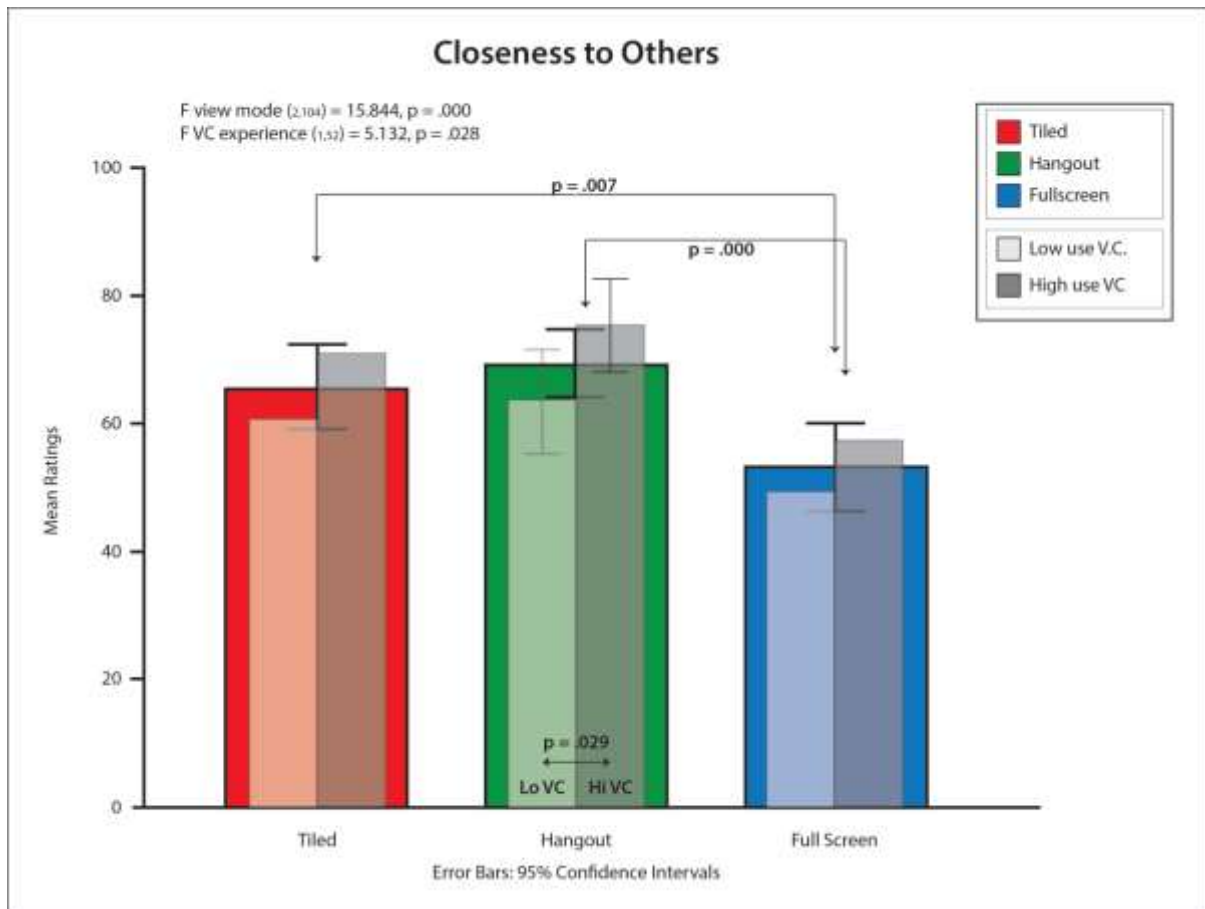


Figure 1.3.2.3.1.5: Closeness to Group

Summary:

The Full Screen view mode detracts significantly from group closeness compared to the other two view modes. In addition, experience with VC significantly enhances group closeness, in particular for the Hangout view mode. The Hangout view mode seems to promote closeness to the group better; at least there is more concordance in the group. The Hangout view mode correlates significantly with both the Tiled and the Full Screen view modes.

Detailed analysis:

The results for a two way ANOVA (view mode * VC experience) are detailed in the tables below.

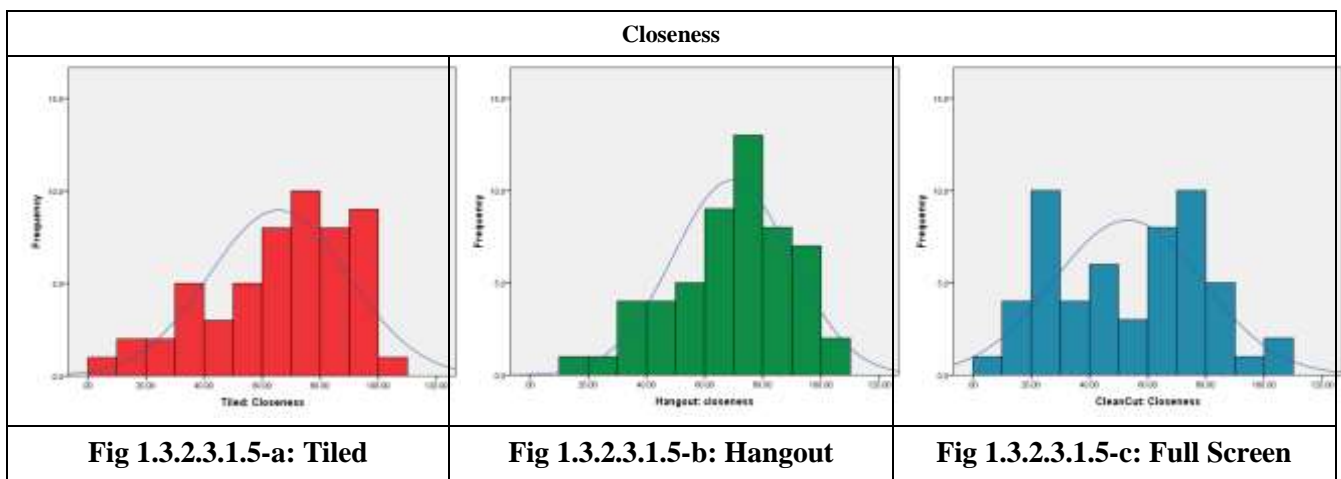
Table 1.3.2.3.1.5a. ANOVA: See who is talking

| Factor | F-ratio | p | η_p^2 |
|-------------------------------|-----------------------|------|------------|
| View Mode | $F_{(2,104)} = 9.874$ | .000 | .160 |
| VC experience | $F_{(1,52)} = 5.132$ | .028 | .090 |
| Follow Up (1 WAY) | $F_{(1,53)}$ | | |
| Tiled vs Hangout | | n.s. | |
| Tiled vs Full Screen | 7.890 | .007 | |
| Hangout vs Full Screen | 23.213 | .000 | |
| Hangout VC experience | 5.074 | .029 | |

Table 1.3.2.3.1.5.b.: Means Seeing others

| | Mean | SD | Lo VC | Hi VC |
|--------------------|-------|-------|-------|-------|
| Tiled | 65.71 | 24.05 | | |
| Hangout | 69.39 | 20.43 | 63.36 | 75.43 |
| Full Screen | 53.21 | 25.67 | | |

The distributions below (graphs 1.3.2.3.1.5-a to c) show some interesting subtle differences. It seems clear from the peak and the low SD (and the highest mean) in the Hangout histogram that this lay-out seems to promote closeness to the group better; at least there is more concordance in the group. The Tiled distribution flattens slightly towards the low end. The Full Screen distribution is more divided in relatively low and high scorers.

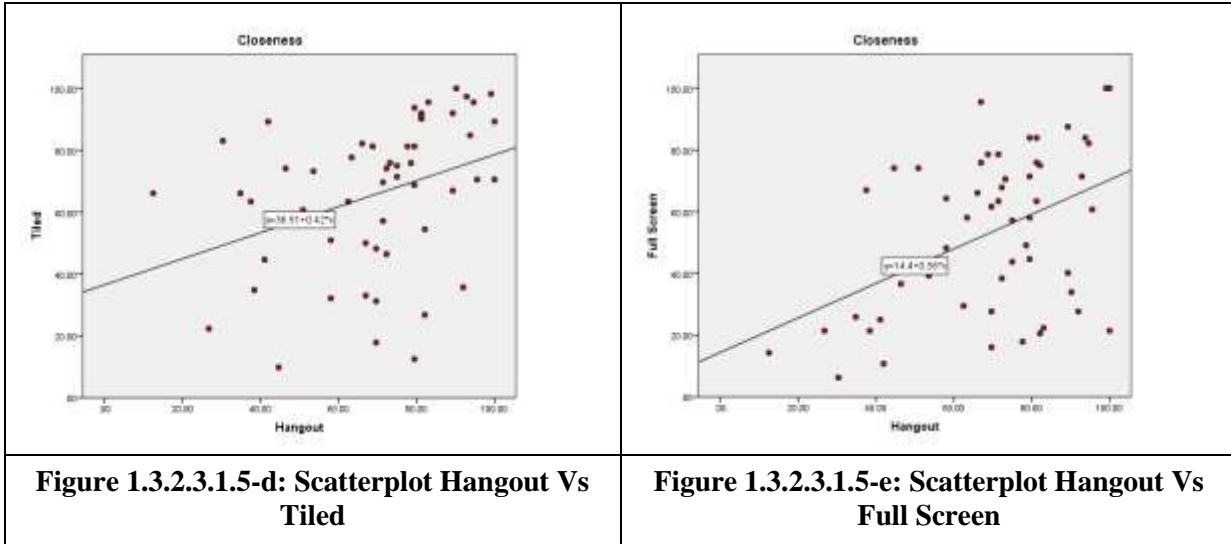


Correlations

The Hangout view mode correlates significantly with the Tiled and the Full Screen view modes.

Table 1.3.2.3.1.5.c: Correlations: Closeness

| | | Tiled | Hangout |
|-------------|-------------|-------------|-------------|
| Hangout | r (df = 53) | .357 | 1 |
| | p | .008 | |
| Full Screen | r (df = 53) | .136 | .445 |
| | p | n.s. | .001 |



- Question 6: How well did see the facial expressions of other people?

Not at all

Very

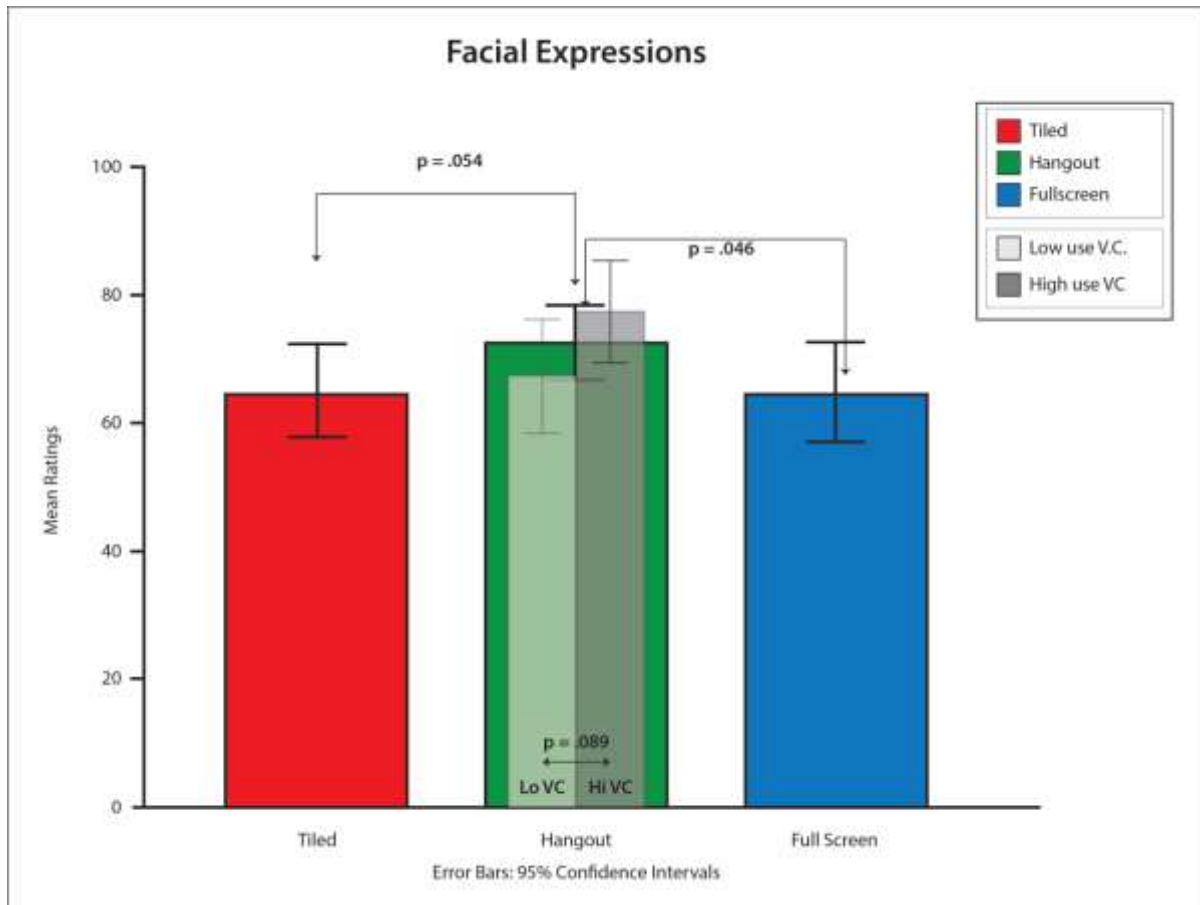


Figure 1.3.2.3.1.6: Seeing facial expressions

Summary:

By and large, there were no differences between the three view modes in the way participants reported how well they could see facial expressions, although the mean for the Hangout view mode was highest and the Full Screen view mode was least. A closer analysis revealed a small but significant difference between those two view modes. For the Full Screen view mode this might have come about because, although there was a peak towards the high end indicating that a group of participants rated this question high as they were able to see the facial expressions of the “speaker” very well, others rated it low, as they could not see the facial expressions of the non-speakers at all. In the Tiled mode the size of the tiles restricted seeing details of facial expressions. The Hangout view mode proved to be a good intermediate solution as it showed facial expressions of the speaker well and participants could still see the rest of the group. There was a trend for high VC users in the Hangout condition to report that they saw facial expressions better than the low VC users. In addition the Hangout view mode correlated significantly with both Full Screen and Tiled viewmodes.

Detailed analysis:

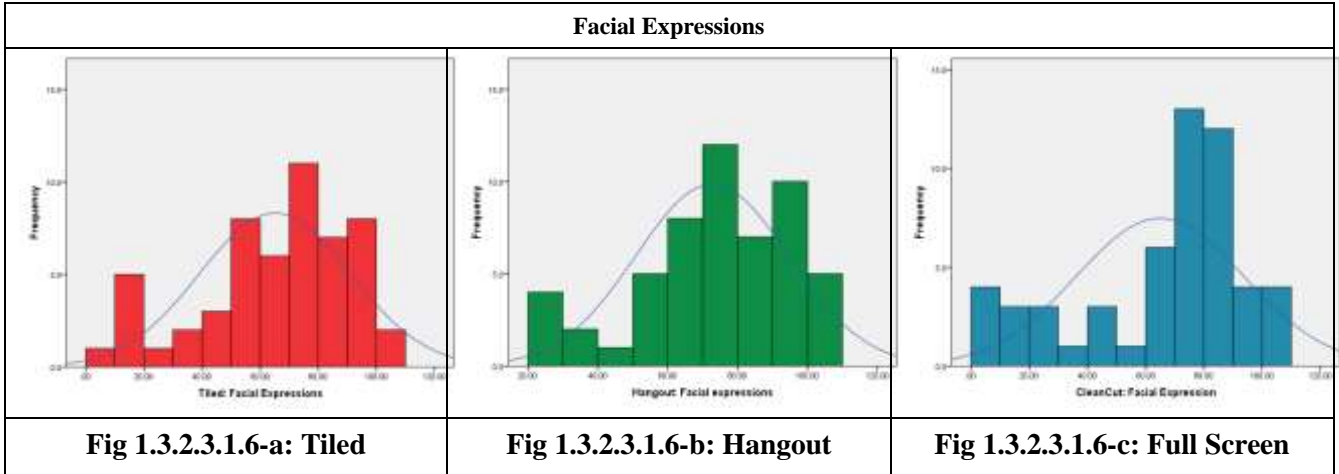
Based on exploratory one way ANOVAs, a two-way ANOVA (view mode * VC experience) was performed, but there were no significant effects. Follow up analysis showed a significant difference between the Hangout and Full Screen view modes.

Table 1.3.2.3.1.6a. ANOVA: Facial Expressions

| Factor | F-ratio | p |
|-------------------------------|--------------|------|
| View Mode | | n.s. |
| Follow Up (1 WAY) | $F_{(1,53)}$ | |
| Tiled vs Hangout | 3.894 | .054 |
| Tiled vs Full Screen | | n.s. |
| Hangout vs Full Screen | 4.165 | .046 |
| Hangout VC experience | 3.000 | .089 |

Table 1.3.2.3.1.6b. Means: Facial Expressions

| | Mean | SD | Lo VC | Hi VC |
|--------------------|-------|-------|-------|-------|
| Tiled | 65.06 | 25.92 | | |
| Hangout | 72.71 | 21.86 | 67.66 | 77.78 |
| Full Screen | 64.91 | 28.82 | | |



Correlations

Table 1.3.2.3.1.6-d: Correlations: Facial Expressions

| | | Tiled | Hangout |
|-------------|-------------|-------------|-------------|
| Hangout | r (df = 53) | .298 | 1 |
| | p | .029 | |
| Full Screen | r (df = 53) | .010 | .412 |
| | p | n.s. | .002 |

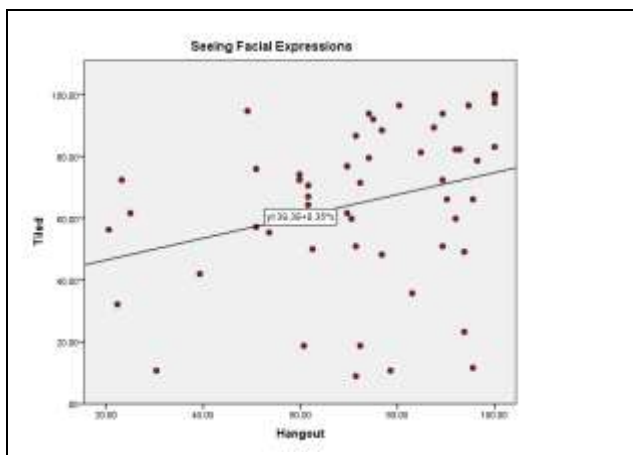


Figure 1.3.2.3.1.6-d: Scatterplot Hangout Vs Tiled

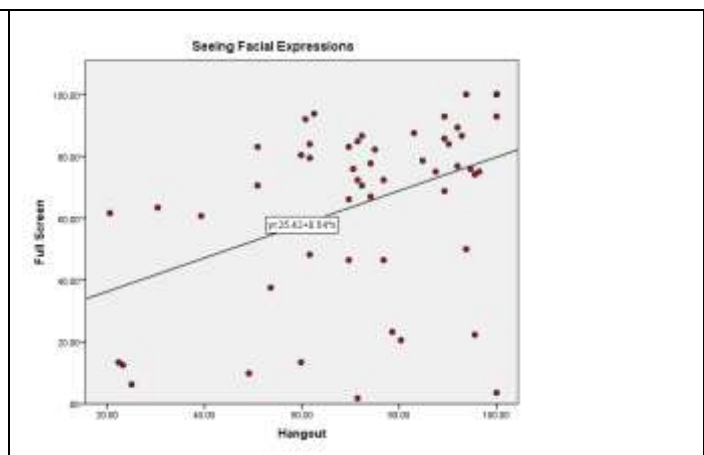


Figure 1.3.2.3.1.6-e: Scatterplot Hangout Vs Full Screen

Question 7: How often did it happen that you and someone else started talking at the same time?

Never

Very



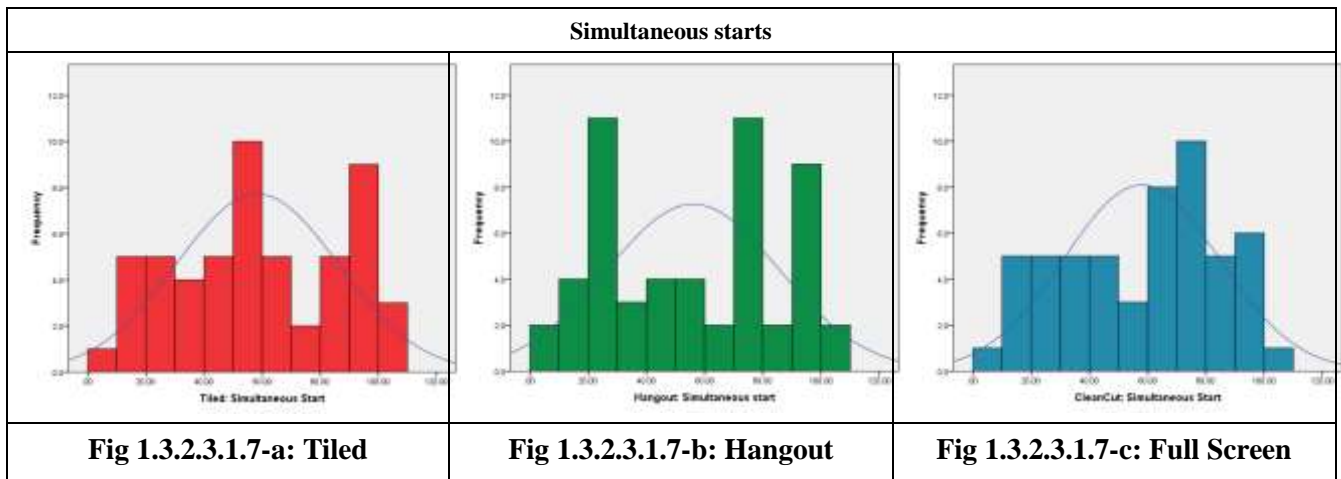
Summary:

There were no significant differences for simultaneous starts between the view modes the means were more than the half way mark, somewhere between 55 and 60, as such simultaneous starts happened relatively often. Compared to the previous questions, here, for all three view modes, there were highly significant inter-correlations, suggesting that independent of view mode participants rated the simultaneous starts in almost identical fashion. It is unclear whether this was affected by delay or signified a healthy conversational flow.

More details

Table 1.3.2.3.1.7b. Means: Simultaneous starts

| | Mean | Standard Deviation |
|-----------------|-------|--------------------|
| Tiled | 58.34 | 27.93 |
| Hangout | 56.01 | 29.68 |
| CleanCut | 57.92 | 26.62 |



Correlations

Table 1.3.2.3.1.7c: Correlations: Simultaneous Starts

| | | Tiled | Hangout |
|-------------|-------------|-------------|-------------|
| Hangout | r (df = 53) | .535 | 1 |
| | p | .000 | |
| Full Screen | r (df = 53) | .522 | .641 |
| | p | .000 | .000 |

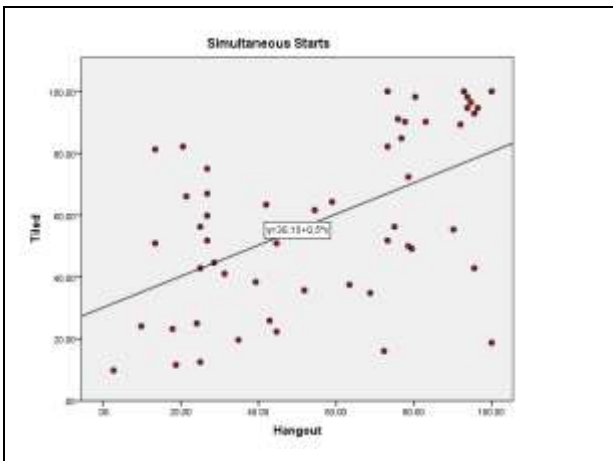


Figure 1.3.2.3.1.7-d: Scatterplot Hangout Vs Tiled

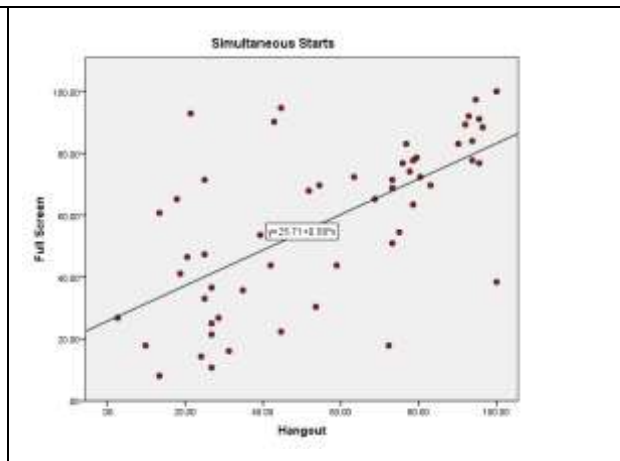


Figure 1.3.2.3.1.7-e: Scatterplot Hangout Vs Full Screen

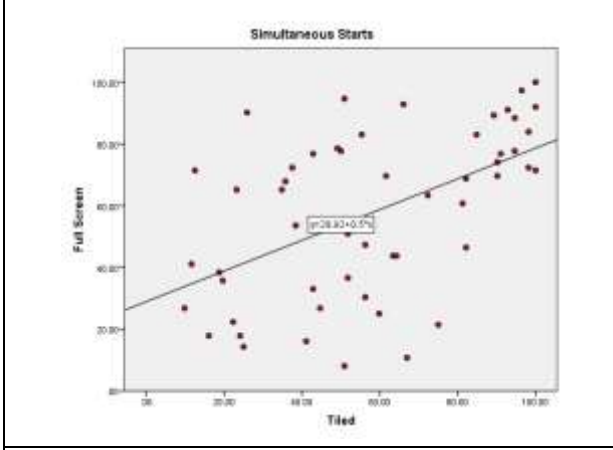
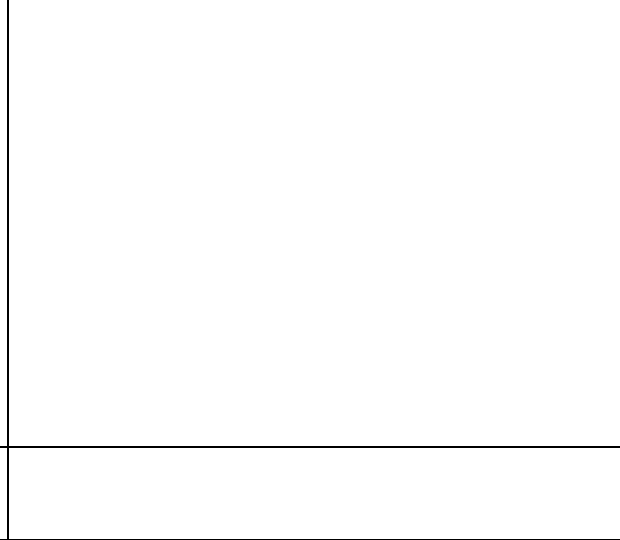


Figure 1.3.2.3.1.7-f: Scatterplot Tiled Vs Full Screen



- Question 8: How often were there awkward silences?

Never Very often

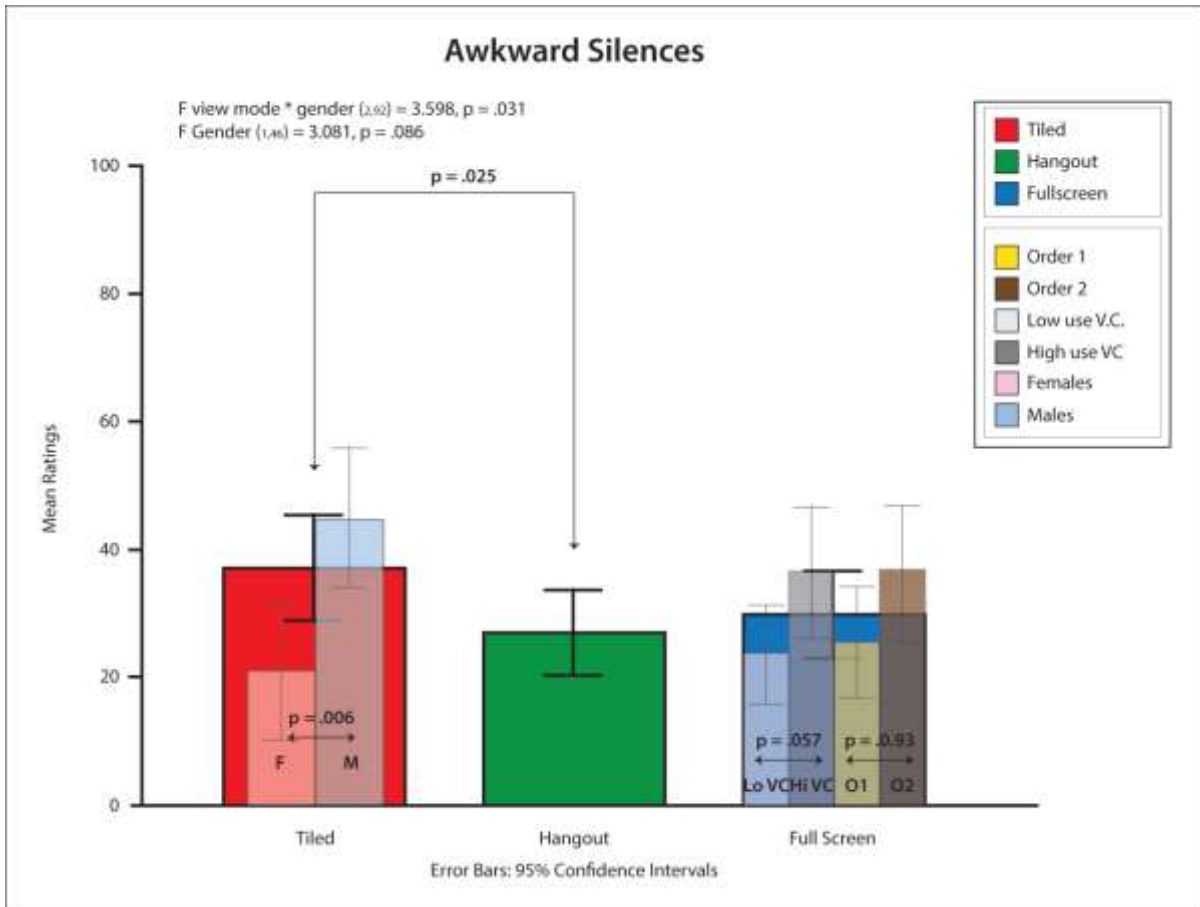


Figure 1.3.2.3.1.8: Awkward Silences

Summary:

There was a low awareness of awkward silences and overall there were no significant differences between the view modes. The mean for the tiled mode was the highest and for the Hangout view mode this was the lowest, this resulted in a significant differences. In the Tiled view mode there was a highly significant gender difference where the males gave much higher ratings than the females. In the Full Screen view mode more experienced high VC users were more aware of the silences and those who started in the Full Screen view mode (Order 2) were more aware than the participants who did this condition last. The Hangout view mode correlated significantly with both the Full Screen and Tilde view modes.

Detailed Analysis:

Based on the exploratory one way ANOVAs it was decided to carry out a four way ANOVA (View Mode * order * gender * VC experience).

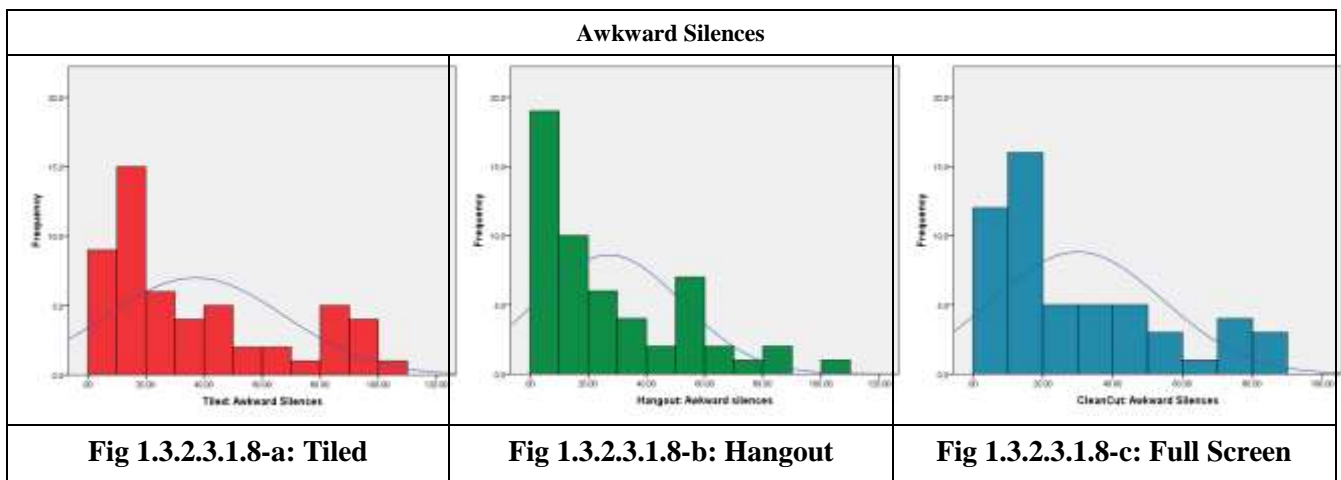
There only was a significant interaction View Mode * Gender and a trend for Gender.

Table 1.3.2.3.1.8a. ANOVA: See who is talking

| Factor | F-ratio | p | η_p^2 |
|---------------------------|--------------------------------|------|------------|
| View Mode | | n.s. | |
| View Mode * Gender | $F_{(2,92)} = 3.598$ | .031 | .073 |
| Gender | $F_{(1,46)} = 3.081$ | .086 | .063 |
| Four way interaction | $F_{(2,92)} = 2.404$ | .096 | .050 |
| Follow Up (1 WAY) | $F_{(1,53)}$ | | |
| Tiled vs Hangout | 5.317 | .025 | |
| Tiled vs Full Screen | | n.s. | |
| Hangout vs Full Screen | | n.s. | |
| Tiled Gender | 8.331 | .006 | |
| Full Screen Order | 2.933 | .093 | |
| Full Screen VC experience | 3.777 | .057 | |

Table 1.3.2.3.1.8b. Means: Awkward Silences

| | Mean | SD | F | M | O1 | O2 | low VC | high VC |
|-------------|-------|-------|-------|-------|-------|-------|--------|---------|
| Tiled | 36.97 | 30.81 | 20.93 | 44.99 | | | | |
| Hangout | 27.08 | 25.05 | | | | | | |
| Full Screen | 29.94 | 24.43 | | | 24.94 | 36.20 | 23.64 | 36.24 |



Correlations

Table 1.3.2.3: Correlations: Awkward Silences

| | | Tiled | Hangout |
|-------------|-------------|-------------|-------------|
| Hangout | r (df = 53) | .378 | 1 |
| | p | .005 | |
| Full Screen | r (df = 53) | .251 | .420 |
| | p | .068 | .002 |

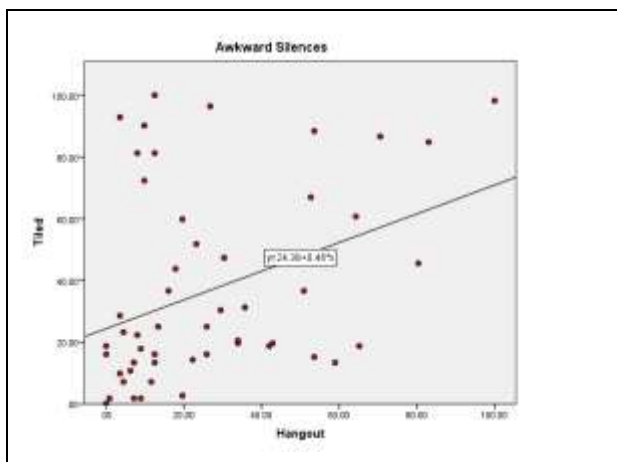


Figure 1.3.2.3.1.8-d: Scatterplot Hangout Vs Tiled

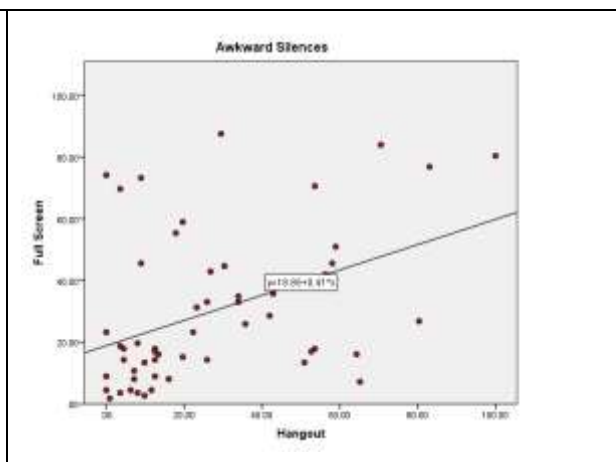


Figure 1.3.2.3.1.8-e: Scatterplot Hangout Vs Full Screen

- Question 9: How lively were the discussions?

Not at all

Very



Summary:

The sessions were very lively and this was reported equally so for each of the view modes. There was a trend in the Hangout view mode for experienced video conferencing users, to find the sessions more lively.

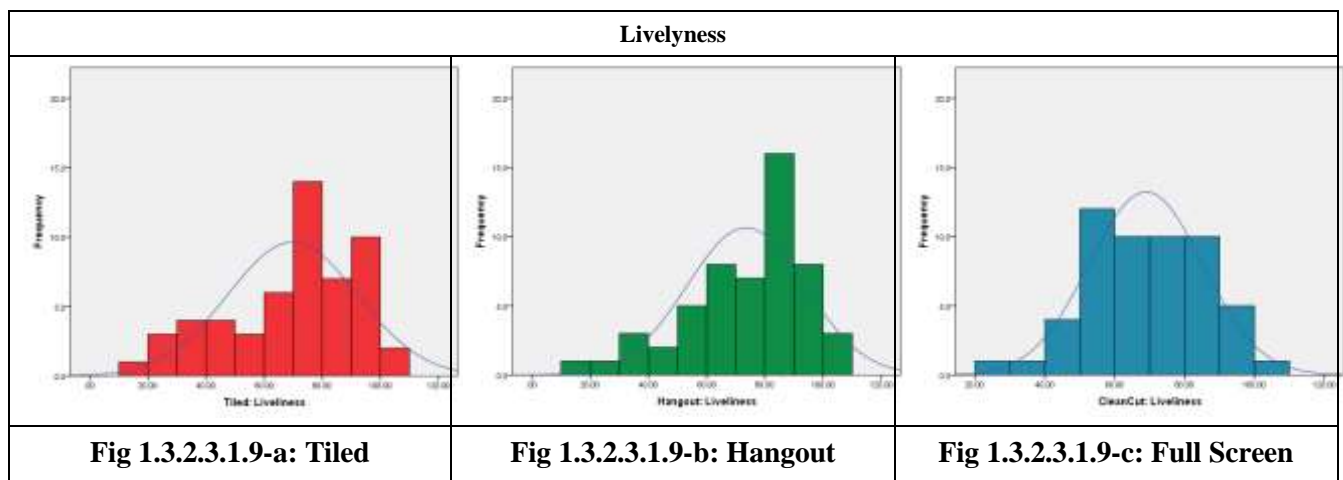


Table 1.3.2.3.1.9: ANOVA's Liveliness

| | Mean | Standard Deviation | VC | VC low | VC high |
|-----------------|-------|--------------------|--------------------------|--------|---------|
| Tiled | 69.77 | 22.35 | | | |
| Hangout | 73.42 | 20.24 | F(1,53) = 3.710 p = .060 | 68.25 | 78.60 |
| CleanCut | 68.94 | 16.28 | | | |

Table 1.3.2.3.1.9b: Correlations: Liveliness

| | | Tiled | Hangout |
|-------------|-------------|-------------|-------------|
| Hangout | r (df = 53) | .515 | 1 |
| | p | .000 | |
| Full Screen | r (df = 53) | .350 | .391 |
| | p | .009 | .003 |

- Question 10: How easy was it to contribute to the discussion?

Not at all

Very

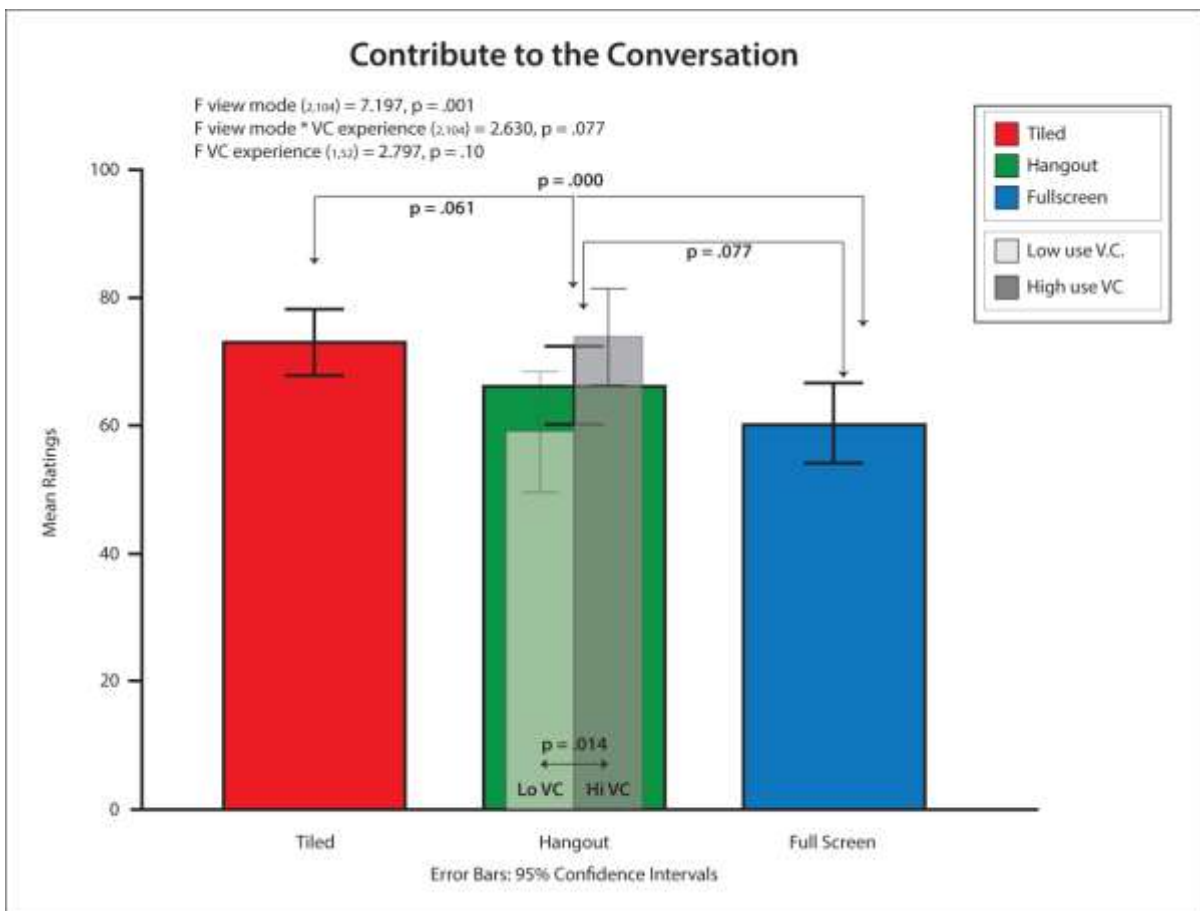


Figure 1.3.2.3.1.10. Contributing to the Conversation

Summary:

The Tiled view mode, i.e. being able to see all the participants equally sized and maintaining position, makes it substantially easier to contribute to the conversation, The Full Screen view mode is significantly worse. In the Hangout view mode, which takes up an intermediary position, participants with more experience in VC find it significantly easier to add to the conversation. The correlations indicate that those who are good (or conversely bad) at making contributions in the Tiled and Hangout view modes are able to cope better in the Full Screen view mode.

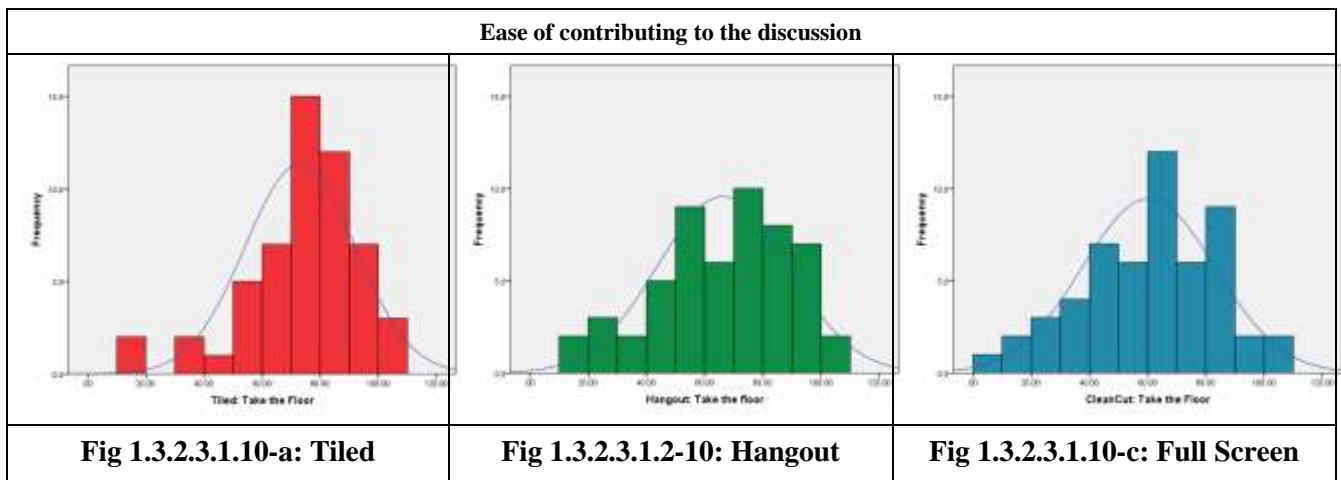
Detailed Analysis

Table 1.3.2.3.1.10a. ANOVA: Take Floor

| Factor | F-ratio | p | η_p^2 |
|---------------------------|-----------------------|------|------------|
| View Mode | $F_{(2,104)} = 7.197$ | .001 | .122 |
| View Mode * VC experience | $F_{(2,104)} = 2.630$ | .077 | .048 |
| VC experience | $F_{(1,52)} = 2.797$ | .10 | .023 |
| Follow Up (1 WAY) | $F_{(1,53)}$ | | |
| Tiled vs Hangout | 3.665 | .061 | |
| Tiled vs Full Screen | 14.174 | .000 | |
| Hangout vs Full Screen | 3.250 | .077 | |
| Hangout VC experience | 6.538 | .014 | |

Table 1.3.2.3.1.10b. Means: Take Floor

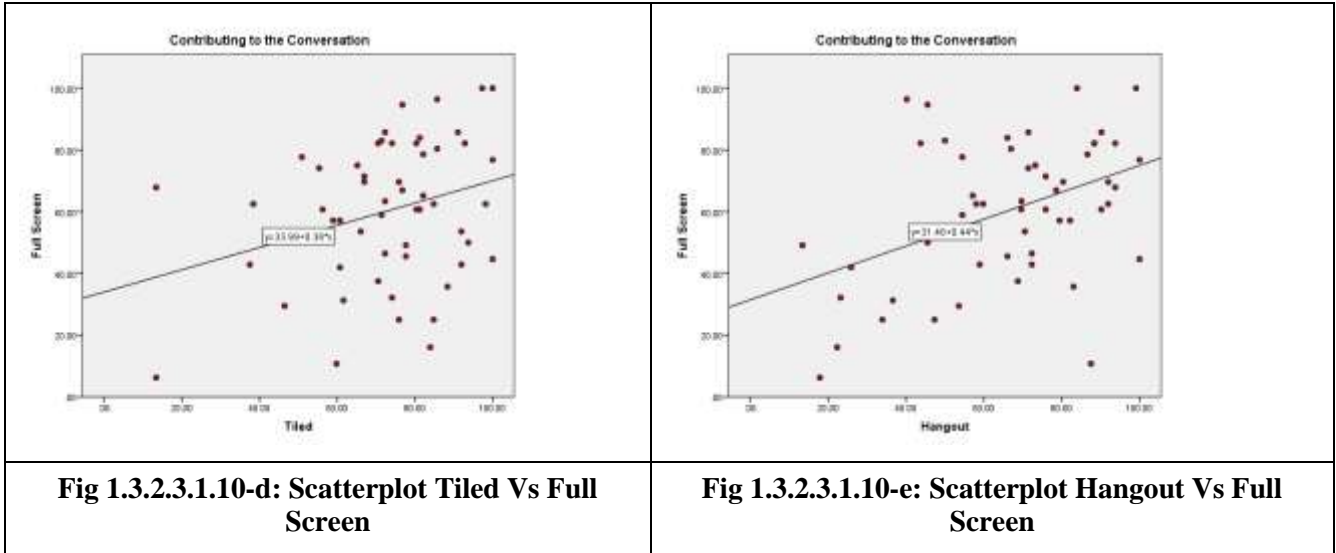
| | Mean | SD | low VC | high VC |
|-------------|-------|-------|--------|---------|
| Tiled | 73.08 | 18.98 | | |
| Hangout | 66.25 | 22.52 | 58.80 | 73.71 |
| Full Screen | 60.31 | 22.81 | | |



Correlations

Table 1.3.2.3.1.10c: Correlations: taking the floor

| | | Tiled | Hangout |
|-------------|-------------|-------------|-------------|
| Hangout | r (df = 53) | .211 | 1 |
| | p | n.s. | |
| Full Screen | r (df = 53) | .300 | .430 |
| | p | .028 | .001 |



Knowing other participants

Before an experimental session started participants were asked how many of the other participants they knew.

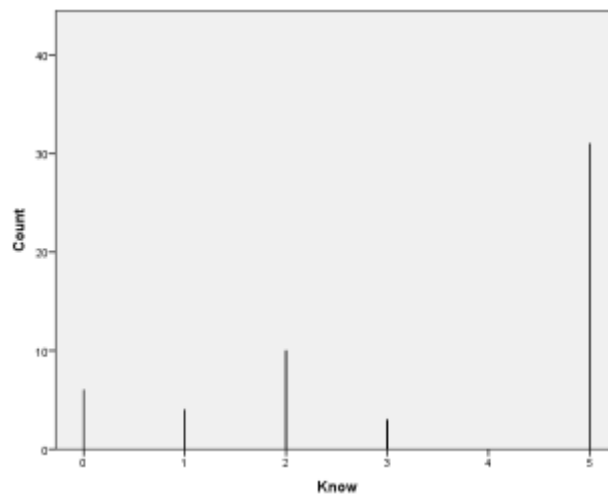


Figure 1.3.2.3.1.11: Knowing other participants

31 participants knew all five of their conversational partners, three knew three of them, as such a total of 34 (out of 54) knew most, if not all, of the other participants.

Six participants knew no one else, four knew one other and 10 knew two participants.

Using this division we derived two categories, know most (,i.e. knew 3 or 5 participants, N = 34) and know few or none (knew 2 or less, N = 20).

We carried out four series of ANOVAs. There were no significant differences for the intensity of using video conferencing or social networks (as might be expected).

There were two near significant differences (trends) in the Tiled view mode, indicating that this design worked as well for those who knew few as those who knew many. In the Tiled view mode there were indications that knowing most of the participants meant it was easier to see others as well as finding them more lifelike.

Table 1.3.2.3.11: Tiled View Mode Knowing many Vs. few

| | F _(1,53) | p | Mean Know | SD Know | Mean Know Few | SD Know Few |
|------------|---------------------|------|-----------|---------|---------------|-------------|
| See others | 3.900 | .054 | 76.76 | 22.68 | 62.41 | 30.43 |
| Lifelike | 3.027 | .088 | 79.39 | 14.90 | 69.69 | 26.17 |

For the Hangout view mode this factor of knowing others resulted in four significant differences and one nearly significant difference. The results however, are in the opposite direction than could be expected. Those who knew few were better able to keep track of the conversation, see other participants, they felt closer to others, experienced fewer simultaneous starts and they found it easier to contribute.

It is not obvious why this happened, but the Hangout view mode worked better for those who knew few than for those who knew each other. As such Hangout accommodates people who know fewer people well.

Table 1.3.2.3.12: Hangout View Mode Knowing many Vs. few

| | F _(1,53) | p | Mean Know | SD Know | Mean Know Few | SD Know Few |
|------------|---------------------|------|-----------|---------|---------------|-------------|
| Keep Track | 11.532 | .001 | 58.82 | 22.19 | 78.39 | 17.01 |
| See others | 5.992 | .018 | 64.68 | 25.11 | 79.96 | 15.70 |
| Close | 3.763 | .058 | 65.36 | 21.90 | 76.25 | 15.90 |
| Sim Starts | 5.932 | .018 | 63.24 | 29.75 | 43.75 | 25.86 |
| Contribute | 10.482 | .002 | 59.24 | 22.51 | 78.17 | 17.25 |

In the Full Screen view mode, those who knew few or none also experienced fewer simultaneous starts but perceived more awkward silences and they found the sessions less lively.

Table 1.3.2.3.13: Full Screen View Mode Knowing many Vs. few

| | F _(1,53) | p | Mean Know | SD Know | Mean Know Few | SD Know Few |
|------------|---------------------|------|-----------|---------|---------------|-------------|
| Sim Starts | 6.312 | .015 | 64.57 | 26.25 | 46.61 | 23.79 |
| Silence | 3.404 | .071 | 25.34 | 24.08 | 37.77 | 23.58 |
| Lively | 11.913 | .001 | 74.29 | 16.48 | 59.87 | 11.42 |

1.3.2.3.2 Discriminant Analysis

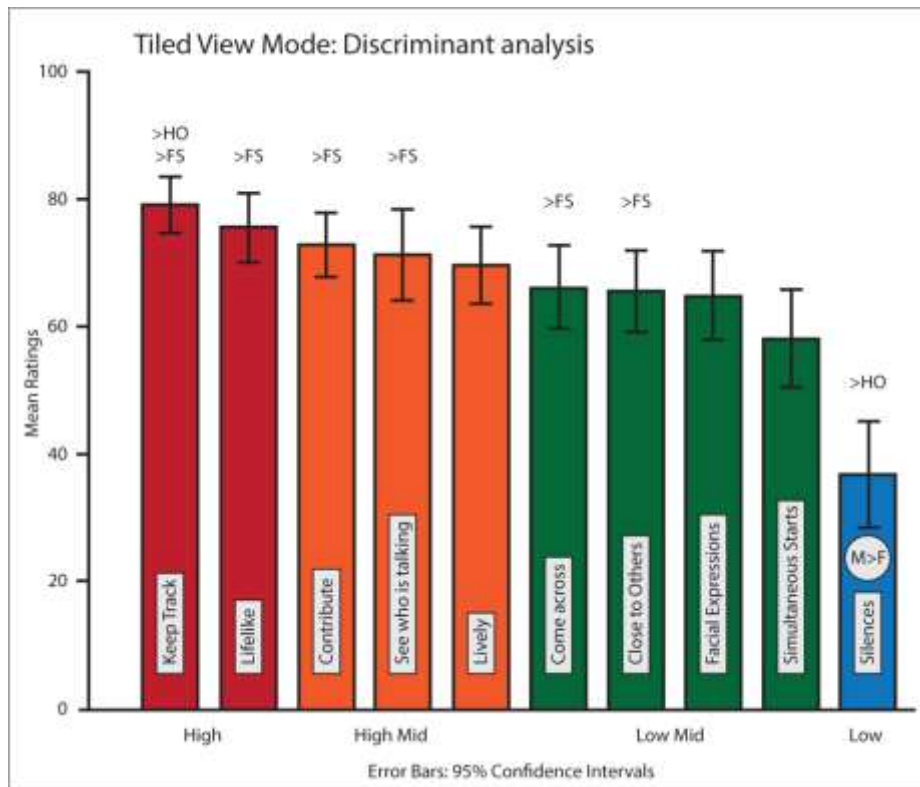


Figure 1.3.2.3.2.1. Tiled View Mode means in descending order

Summary Tile View Mode:

We were able to divide the ratings for the Tiled view mode into four separate bands: High, high mid, low mid and low ratings.

1. High Ratings (Means 76 – 79)

The two highest means of the questionnaire for all three view modes are for how in the Tiled View mode easy it was to Keep Track of the conversation, significantly higher than the Hangout and Full Screen view modes, and how Lifelike participants were significantly higher than the Full Screen view mode.

The first finding can be understood from the existing literature pointing out how important it is to be able to see all participants at all times in a consistent screen location. The second finding comes as a bit of a surprise given the smaller tile sizes for the Tiled view mode but is most likely a confirmation that if the display is consistent and all participants are visible at all times this also results in conversational partners seeming lifelike.

2. High Mid Ratings (Means 70 – 73)

The band of high mid ratings consists of the ease of Contributing to the conversation and being able to See who is Talking, both significantly higher than the Full Screen ratings. The rationale that being able to monitor the others, including being able to see who is talking, at all time in a consistent location on the screen makes it easier to contribute to the conversation. The nature of the task results in very lively conversations, and here there are no differences with the other two view modes. Irrespective of how well a view mode supported mediated group communication, the conversations were lively.

3. Low Mid Ratings (Means 58 – 66)

This grouping is somewhat marred by arbitrariness, as feeling how well participants were Coming Across did not significantly differ from being able to see who is talking and how lively the conversation was. As such Coming Across takes up an intermediary position. Coming Across and feeling Close to Others are both rated significantly than in the Full Screen mode. All the same it shows that the mediated communication enabled the business of talking to each other (keeping track, contributing, seeing who is talking) better than feeling close (like in a face to face situation). Oddly enough, although participants seemed very lifelike, this did not include being able to see facial expressions that well. Simultaneous starts were still rated over the half way mark, but it is unclear to which extent delay played a role here. In order to keep the questionnaire between conditions to an absolute minimum, we omitted questions about the perception of delay.

4. Low Ratings (Mean 37)

As the conversations were lively the occurrence of awkward silences was rated as low. However, for the males this was significantly higher than in the Hangout (HO) condition. It is possible that because the Hangout interface promotes a more formalised turn taking (by those who speak loudest) and the Tiled view mode a more spontaneous democratic group discussion awkward silences occasionally come about more often as the males are hesitant to proceed and fall silent.

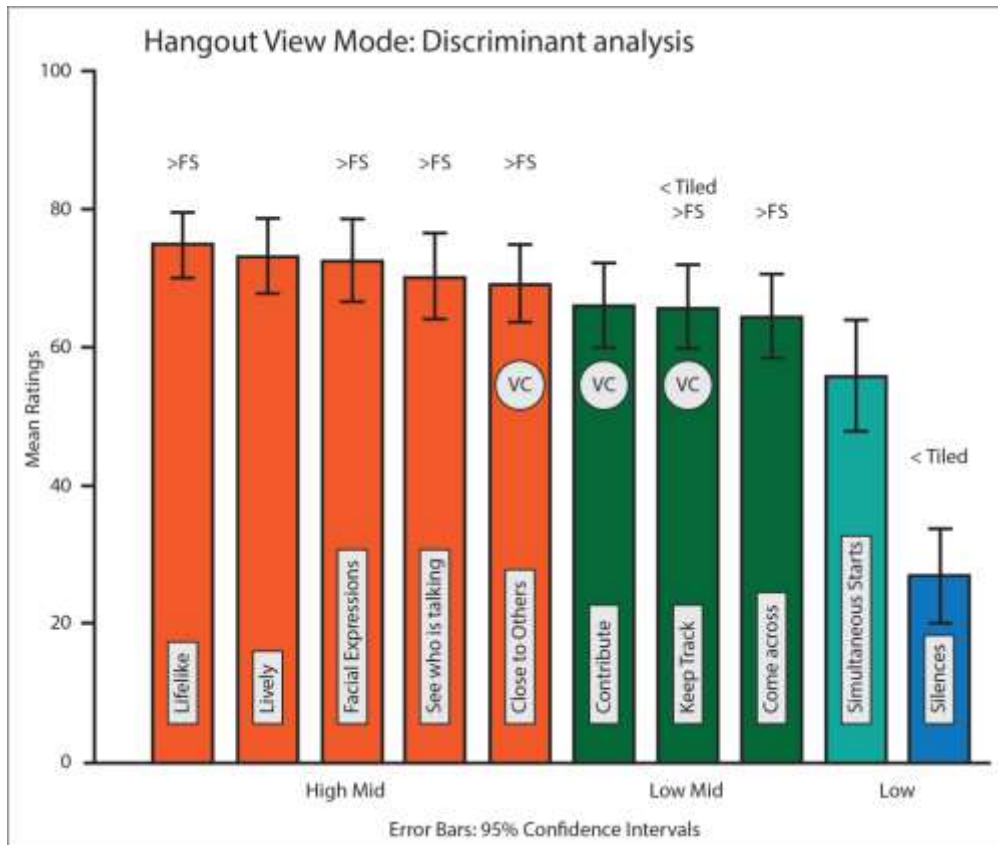


Figure 1.3.2.3.2.2. Hangout view mode, means in descending order

Summary Hangout view mode:

In line with grouping the means in the Tiled conditions and following the same colour coding, here we can distinguish three groups: High Mid, Low Mid and Low. The results may have been affected that the smaller windows (tiles) are so much smaller in comparison to the main window (the speaker window) and in addition as there are only five smaller tiles, there is no consistency of location for the non-speakers.

1. High Mid Ratings (Means 69 – 75)

Although there are no significant differences with the Tiled view modes, there are four variables that are significantly higher than in the Full Screen View Mode: How Lifelike participants appeared, seeing Facial Expressions, seeing Who is Talking and feeling Close to Others. The more experienced Video Conferencing users (VC) feel significantly closer than those with less experience.

2. Low Mid Ratings (Means 65 – 66)

All three items in this band are rated significantly higher than the Full Screen Mode, Contributing to the discussion and being able to Keep Track of the conversation is easier for the experienced VC users. The latter is rated significantly lower than in the Tiled view mode. The two items and, how participants feel they come across is rated significantly higher than the Full Screen view mode.

3. Low Ratings (Means 27 – 56)

Simultaneous starts occur relatively low and there are significantly fewer awkward Silences in the Hangout view mode than in the Tiled one. Because of the significant differences between the two items Simultaneous starts is depicted in a paler blue than the awkward Silences.

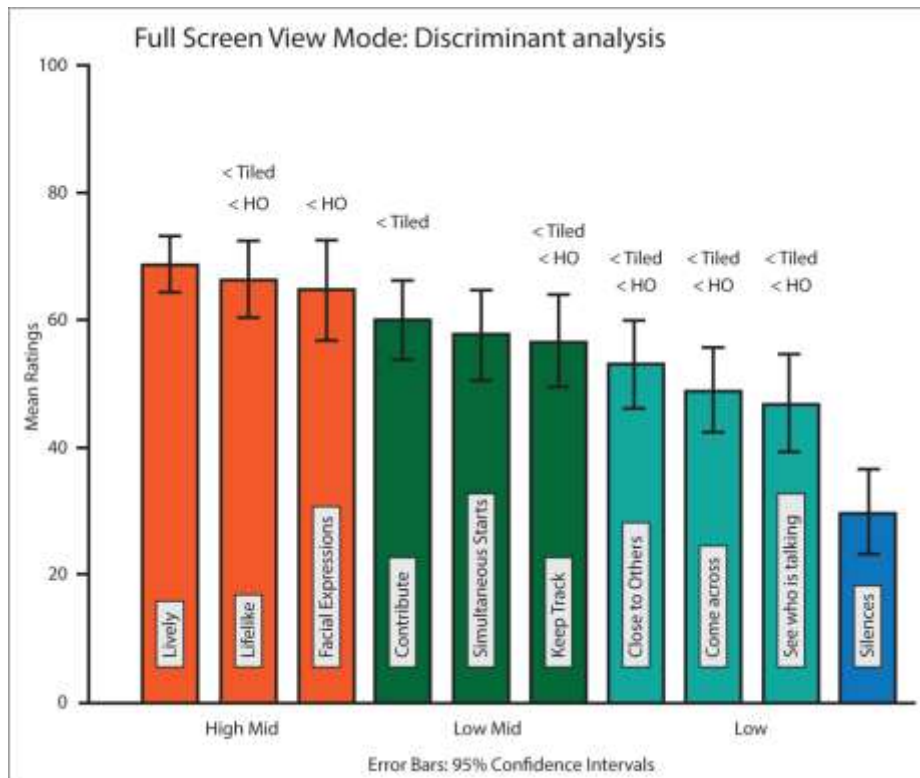


Figure 1.3.2.3.2.3. Full Screen view mode, means in descending order

Summary

Deviating slightly from the guidelines that we used for the previous two view modes, we derive three bands of ratings.

1. High Mid Ratings (Means 65 – 69)

The sessions were lively in all three view mode conditions and in the Full Screen view mode this receives the highest ratings. The size of the view mode seems beneficial for participants being relatively lifelike and being able to see facial expressions.

2. Low Mid Ratings (Means 57 – 60)

The low mid ratings are made of not being able to contribute to the conversation as well as in the Tiled view mode, nor being able to keep track of the conversation. On the plus side, the simultaneous starts are independent of view mode.

3. Low Ratings (Means 30 – 53)

In contrast to the previous two view modes there are three sets of low ratings indicating a lack of closeness to the others, not feeling they come across well to the others, not being able to see who is talking (it is possible that the livingness of the conversation put quite a strain on the Orchestration engine, although the same engine drove the Hangout mode). On a better note, there are few awkward silences.

Detailed analysis

Tables 1.3.2.3.2.1 – 3 show the means, standard error, standard deviations, skewness and kurtosis for the questions in the Tiled, Hangout and Full Screen view modes respectively. The means are in the same order as the bar charts above.

Table 1.3.2.3.2.1: Statistical Descriptions Tiled view mode

| | Mean | Std. Error | SD | Skewness | Kurtosis |
|-------------|---------|------------|----------|----------|----------|
| Keep Track | 79.2824 | 2.20523 | 16.20506 | -1.176 | 1.918 |
| Lifelike | 75.7937 | 2.74256 | 20.15360 | -1.058 | .629 |
| Contribute | 73.0820 | 2.58344 | 18.98431 | -1.214 | 2.143 |
| See others | 71.4451 | 3.60338 | 26.47932 | -.788 | -.634 |
| Lively | 69.7751 | 3.04149 | 22.35032 | -.809 | -.172 |
| Come across | 66.4021 | 3.26711 | 24.00823 | -.884 | .286 |
| Close | 65.7077 | 3.27394 | 24.05844 | -.658 | -.432 |
| Face | 65.0628 | 3.52843 | 25.92859 | -.713 | -.296 |
| Sim Starts | 58.3499 | 3.80130 | 27.93371 | -.035 | -1.183 |
| Silence | 36.9709 | 4.19203 | 30.80498 | .777 | -.730 |

Table 1.3.2.3.2.2: Statistical Descriptions Hangout view mode

| | Mean | Std. Error | SD | Skewness | Kurtosis |
|-------------|---------|------------|----------|----------|----------|
| Lifelike | 74.9835 | 2.37105 | 17.42355 | -.799 | .428 |
| Lively | 73.4292 | 2.75464 | 20.24242 | -1.000 | .503 |
| Face | 72.7183 | 2.97560 | 21.86609 | -.843 | .118 |
| See others | 70.3373 | 3.15215 | 23.16348 | -.759 | -.186 |
| Close | 69.3948 | 2.78026 | 20.43064 | -.718 | .051 |
| Contribute | 66.2533 | 3.06503 | 22.52327 | -.580 | -.387 |
| Keep Track | 66.0714 | 3.04676 | 22.38905 | -.601 | -.137 |
| Come across | 64.6825 | 2.99166 | 21.98409 | -.450 | -.476 |

| | | | | | |
|------------|---------|---------|----------|-------|--------|
| Sim Starts | 56.0185 | 4.03911 | 29.68131 | -.069 | -1.454 |
| Silence | 27.0833 | 3.40894 | 25.05049 | 1.026 | .244 |

Table 1.3.2.3.2.1: Statistical Descriptions Full Screen view mode

| | Mean | Std. Error | SD | Skewness | Kurtosis |
|-------------|---------|------------|----------|----------|----------|
| q309 | 68.9484 | 2.21623 | 16.28588 | -.330 | -.156 |
| Lifelike | 66.6171 | 2.97557 | 21.86592 | -1.011 | .467 |
| Face | 64.9140 | 3.92243 | 28.82387 | -.960 | -.282 |
| Contribute | 60.3175 | 3.10326 | 22.80419 | -.401 | -.378 |
| Sim Starts | 57.9200 | 3.62254 | 26.62013 | -.288 | -1.167 |
| Keep Track | 57.0602 | 3.62340 | 26.62647 | -.031 | -1.234 |
| Close | 53.2077 | 3.49348 | 25.67176 | -.071 | -1.149 |
| Come across | 49.1402 | 3.36519 | 24.72897 | -.039 | -.771 |
| q303 | 47.0899 | 3.88372 | 28.53939 | .436 | -1.154 |
| Silence | 29.9438 | 3.32520 | 24.43511 | .955 | -.211 |

Tables 1.3.2.3.2.4 – 6 show in the same order as the variables in the previous tables and bar charts, the p-values for the paired comparisons exercise on which we based the discriminant analysis.

Table 1.3.2.3.2.4. p-values paired comparisons Tiled view mode

| | track | lifelike | floor | see | lively | across | close | face | sim start |
|-----------|--------------|--------------|--------------|--------------|--------------|----------|----------|----------|--------------|
| lifelike | ns | | | | | | | | |
| floor | 0.008 | ns | | | | | | | |
| see | 0.009 | 0.084 | ns | | | | | | |
| lively | 0 | 0.055 | ns | ns | | | | | |
| across | 0 | 0.005 | 0.014 | ns | ns | | | | |
| close | 0 | 0.001 | 0.017 | ns | ns | ns | | | |
| face | 0 | 0 | 0.021 | 0.035 | ns | ns | ns | | |
| sim start | 0 | 0.001 | 0.003 | 0.02 | 0.008 | ns | ns | ns | |
| silence | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.001 |



Table 1.3.2.3.2.5. p-values paired comparisons Hangout view mode

| | lifelike | lively | face | see | close | floor | track | across | sim start |
|-----------|----------|--------|-------|-------|-------|-------|-------|--------|-----------|
| lively | ns | | | | | | | | |
| face | ns | ns | | | | | | | |
| see | ns | ns | ns | | | | | | |
| close | 0.017 | ns | ns | ns | | | | | |
| floor | 0.008 | 0.027 | 0.033 | 0.043 | ns | | | | |
| track | 0.012 | 0.051 | 0.054 | ns | ns | ns | | | |
| across | 0.001 | 0.007 | 0.005 | 0.059 | 0.055 | ns | ns | | |
| sim start | 0 | 0 | 0 | 0.016 | 0.01 | 0.083 | 0.07 | 0.098 | |
| silence | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 1.3.2.3.2.6. p-values paired comparisons Full Screen view mode

| | lively | lifelike | face | floor | sim start | track | close | across | see |
|-----------|--------|----------|-------|-------|-----------|-------|-------|--------|-------|
| lifelike | ns | | | | | | | | |
| face | ns | ns | | | | | | | |
| floor | 0.011 | 0.075 | ns | | | | | | |
| sim start | 0.002 | 0.08 | ns | ns | | | | | |
| track | 0.005 | 0.024 | ns | ns | ns | | | | |
| close | 0 | 0 | 0.002 | 0.02 | ns | ns | | | |
| across | 0 | 0 | 0 | 0.001 | ns | 0.017 | ns | | |
| see | 0 | 0 | 0 | 0.002 | 0.069 | 0.036 | ns | ns | |
| silence | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.003 |

1.3.2.3.3 Correlations

3.4.3.3.1. Tiled View Mode

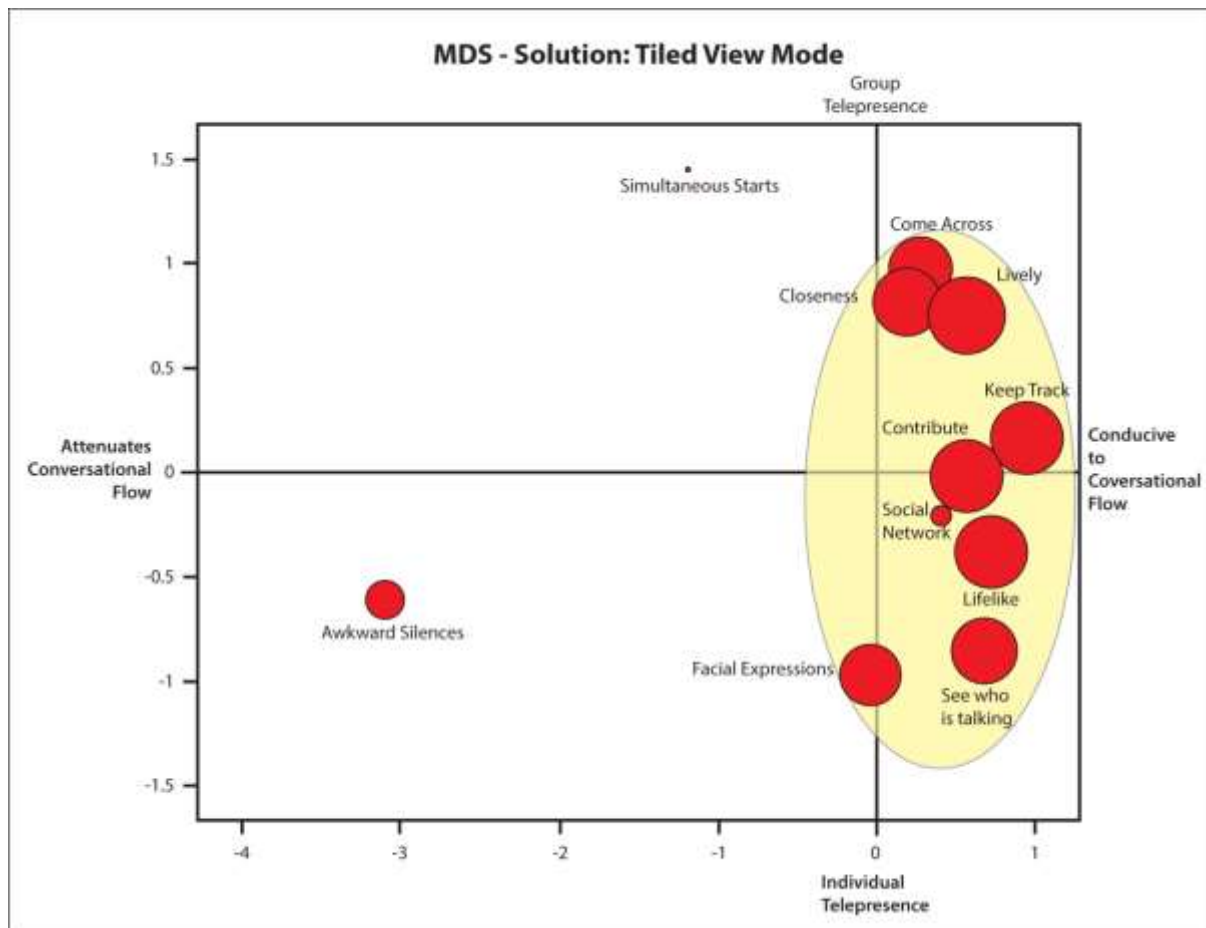


Figure 1.3.2.3.3.1. MDS solution Tiled View Mode

Summary

We derived a correlation matrix for the ratings for the pre-experiment questions on how intensive participants used video conferencing and social networks and the ratings of the questions for the Tiled view mode. The intensity of video conferencing did not correlate significantly with any of the view mode ratings but the intensity of using social networks correlated significantly with the ratings of four of the questions, in particular there was a strong correlation with being able to contribute to the discussion.

In other words those who reported to use social media more intensively (and the mean was a high 77) also found it easier to contribute to the conversation, found the other participants more lifelike, could see better who was talking and saw facial expressions better.

With the exception of Simultaneous Starts which correlated significantly (and positively) only with one other variable (Liveliness), almost all the variables showed a high number of extremely significant inter-correlations and these are reflected in the size of the circles in the plot above (figure 1.3.2.3.3.1., detailed in table 1.3.2.3.3.1.) resulting in one big cluster consisting of nine variables. The occurrence of Awkward Silences is strongly negatively correlated with seven of those this cluster.

The axes were named to reflect high ratings: those who kept better track of the conversation, also contributed more, also found the sessions livelier. However, it is good to bear in mind that the opposite is also true, if you found it more difficult to keep track than it was also more difficult to contribute. Those variables



are conducive to conversational flow, whereas the occurrence of awkward silences (even though they were relatively rare) is a variable that is not conducive to the conversational flow. As such the X-axis reflects conversational flow.

Seeing an individual’s facial expression, see which individual is talking is indicative for the telepresence of an individual. Feeling that you come across well to the group, feeling close to other participants and feeling that the group discussion was lively indicates group telepresence. Therefore the one extreme of Y-axis was interpreted as “individual telepresence” and as other “group telepresence”.

Detailed Analysis

Table 1.3.2.3.3.1.: Tiled View Mode significant correlations

| TILED | SN | Track | Across | See | Lifelike | Close | Face | SimStarts | Silence | Lively | Contribute |
|--------------|----------|----------|----------|----------|----------|----------|----------|-----------|----------|----------|------------|
| SN | | | | | | | | | | | |
| Track | | | | | | | | | | | |
| Across | | .000 | | | | | | | | | |
| See | .044 | .000 | .006 | | | | | | | | |
| Lifelike | .034 | .000 | .001 | .000 | | | | | | | |
| Close | | .000 | .000 | .001 | .000 | | | | | | |
| Face | .070 | .001 | .014 | .000 | .000 | .000 | | | | | |
| SimStarts | | | | | | | | | | | |
| Silence | | .000 | .019 | .071 | .034 | .037 | | | | | |
| Lively | | .000 | .000 | .000 | .001 | .000 | .002 | .036 | .000 | | |
| Contribute | .006 | .000 | .000 | .014 | .004 | .000 | .002 | | .015 | .004 | |
| Total | 4 | 8 | 8 | 9 | 9 | 8 | 8 | 1 | 7 | 9 | 9 |
| p<.01 | 1 | 8 | 6 | 6 | 7 | 7 | 6 | | 2 | 8 | 7 |
| p<.05 | 2 | | 2 | 2 | 2 | 1 | 1 | 1 | 4 | 1 | 2 |
| p<.10 | 1 | | | 1 | | | 1 | | 1 | | |
| Size | 4.5 | 16 | 14 | 14.5 | 16 | 15 | 13.5 | 1 | 8.5 | 17 | 16 |

Table 1.3.2.3.3.1. shows the p-values for the significant correlations, where red indicates negative correlations (we omit the empty column showing that there were no significant correlations with the level of video conferencing use). For each variable the total of significant correlations are shown in the yellow row. The rows below show how many of these were significant at the p<.01, p <.05 and p<.10 level. The last row indicates the size of the circles in figure 1.3.2.3.3.1. based on a weight of “2” given to a significance <.01, a weight of “1” for a significant p-value <.05 and a weight of “0.5” for significance <.10. The numbers in the row labelled “Total” add up to 80, as the matrix contains a total of 40 significant correlations between two variables. SN stands for Social Networking. The other abbreviations refer to the chronological order of the questions as outlined in section 1.3.2.3.1.

3.4.3.3.2. Hangout view mode

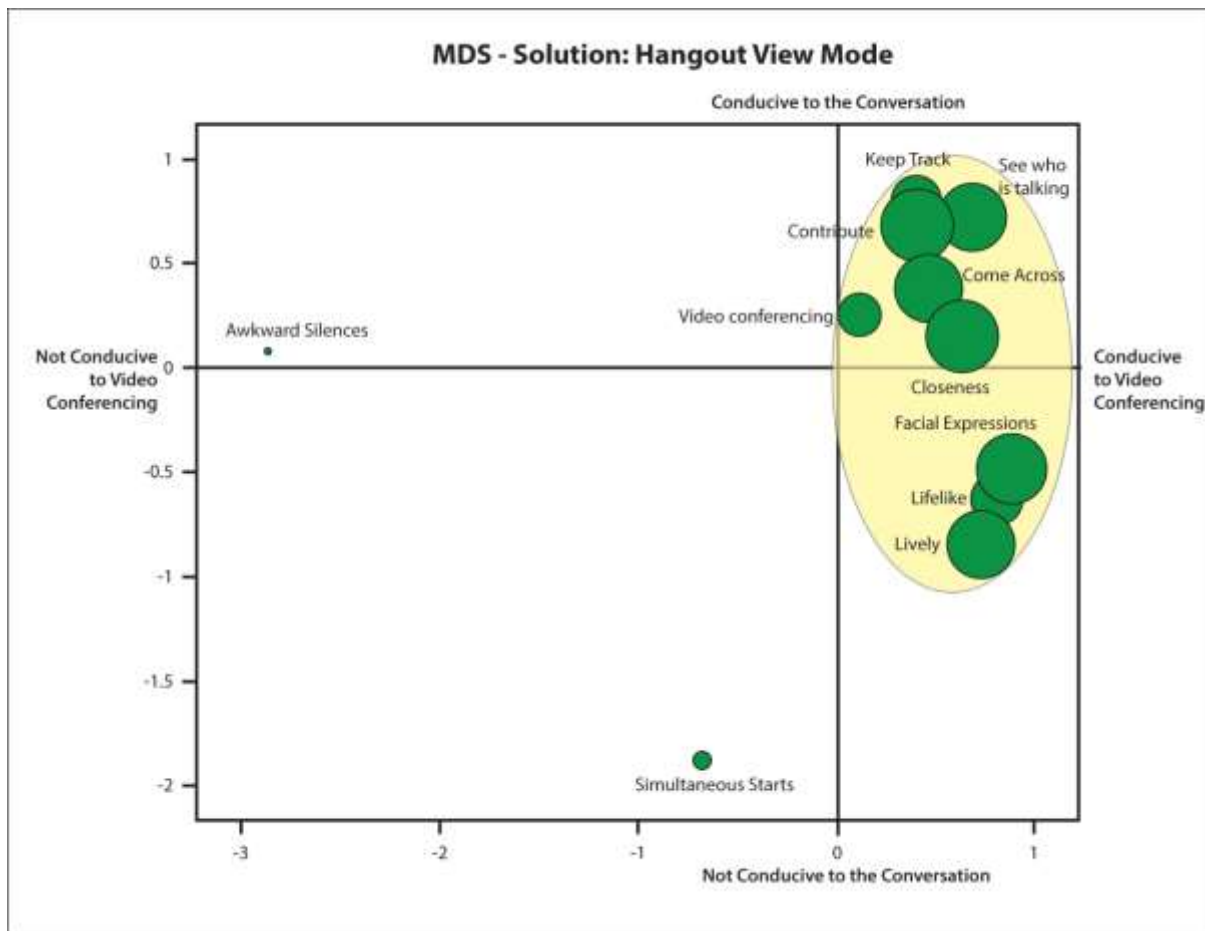


Figure 1.3.2.3.3.2. MDS solution Hangout View Mode

Summary

We derived a correlation matrix for the ratings for the pre-experiment questions on how intensive participants used video conferencing and social networks and the ratings of the questions for the Hangout view mode. In contrast to the Tiled view mode, the intensity of video conferencing correlated significantly with eight of the Hangout view mode ratings but the intensity of using social networks did not yield any significant correlations.

Thus those who have more video conferencing experience, perform better in the Hangout view mode. Here it is good to reiterate that in the like for like comparisons (section 1.3.2.3.1.) there was a distinctive lack of significant correlations between the answers to the (same) questions for the Tiled view mode and Hangout view mode and a proliferation of significant correlations between the Hangout and Full Screen view modes.

This suggests that the experience in the Hangout view mode is quite different from using the Tiled view mode. The Tiled view mode seems to map on to the (intensity of the) use of social networks and the Hangout view mode on the use of video conferencing.

Almost all the variables showed a high number of extremely significant inter-correlations and these are reflected in the size of the circles in the plot above (figure 1.3.2.3.3.2., detailed in table 1.3.2.3.3.2.) resulting in one big cluster consisting of nine variables. In contrast to the Tiled view mode the occurrence of Simultaneous Starts is significantly correlated with three items in the main cluster, i.e. a strong positive



correlation with Liveliness and two negative correlations with being able to contribute to the discussion and with being able to see other participants.

Here it was less straight forward to name the X and Y axes. In fact the main cluster represents aspects of video conferencing without much differentiating. Tentatively the X-axis is interpreted as representing video conferencing ability and the Y-axis representing aspects that were conducive to talking about the subject matter.

It seems tempting to paraphrase that the Tiled view mode emphasises socialising interactively with people, possibly an explanation as to why in particular females prefer this view mode, whereas the Hangout view mode emphasis talking about “things” [Whitaker, 199?] where previous experience with video conferencing is of importance.

Detailed Analysis

Table 1.3.2.3.3.2.: Hangout View Mode significant correlations

| HANGOUT | VC | Track | Across | See | Lifelike | Close | Face | SimStarts | Silence | Lively | Contribute |
|----------------|----------|----------|----------|----------|----------|----------|----------|-----------|----------|----------|------------|
| VC | df=53 | | | | | | | | | | |
| Track | .016 | | | | | | | | | | |
| Across | .028 | .000 | | | | | | | | | |
| See | .040 | .005 | .000 | | | | | | | | |
| Lifelike | .061 | | .001 | .005 | | | | | | | |
| Close | .004 | .000 | .000 | .000 | .000 | | | | | | |
| Face | .028 | .006 | .000 | .000 | .001 | .000 | | | | | |
| SimStarts | | | | .048 | | | | | | | |
| Silence | | | | | | | .086 | | | | |
| Lively | .012 | | .002 | .048 | .003 | .003 | .000 | .002 | .017 | | |
| Contribute | .004 | .000 | .000 | .000 | .014 | .000 | .000 | .021 | | .002 | |
| Total | 8 | 6 | 8 | 9 | 7 | 8 | 9 | 3 | 2 | 9 | 9 |
| p<.01 | 2 | 5 | 7 | 6 | 5 | 8 | 7 | 1 | | 6 | 7 |
| p<.05 | 5 | 1 | 1 | 3 | 1 | | 1 | 2 | 1 | 3 | 2 |
| p<.10 | 1 | | | | 1 | | 1 | | 1 | | |
| Size | 9.5 | 11 | 15 | 15 | 11.5 | 16 | 15.5 | 4 | 1.5 | 15 | 16 |

Table 1.3.2.3.3.2. shows the p-values for the significant correlations , where red indicates negative correlations (we omit the empty column showing that there were no significant correlations with the level of social network use). For each variable the total of significant correlations are shown in the yellow row. The rows below show how many of these were significant at the p<.01, p <.05 and p<.10 level. The last row indicates the size of the circles in figure 1.3.2.3.3.2. based on a weight of “2” given to a significance <.01, a weight of “1” for a significant p-value <.05 and a weight of “0.5” for significance <.10. The numbers in the row labelled “Total” add up to 78, as the matrix contains a total of 39 significant correlations between two variables. VC stands for Video Conferencing. The other abbreviations refer to the chronological order of the questions as outlined in section 1.3.2.3.1.

3.4.3.3.3. Full Screen View Mode

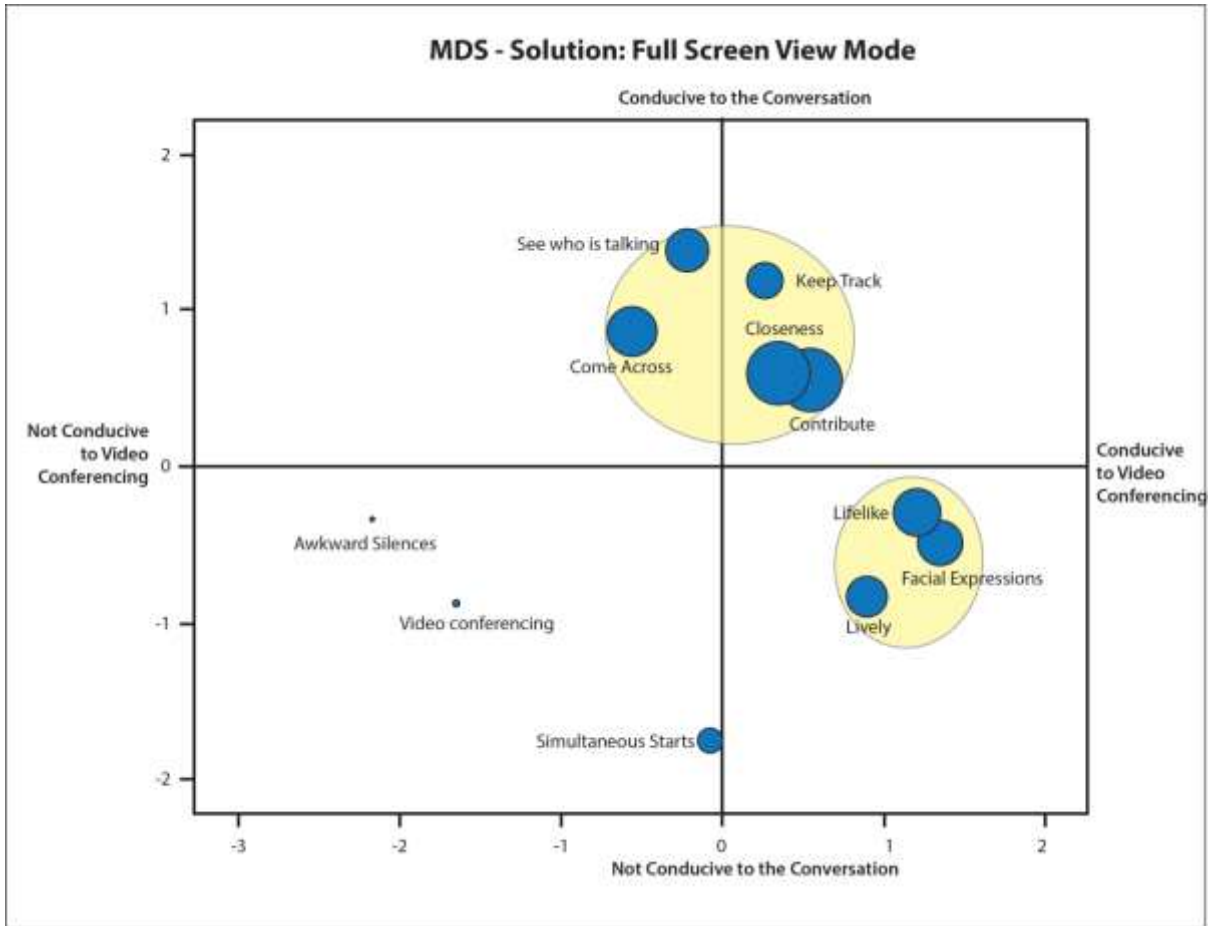


Figure 1.3.2.3.3.3. MDS solution Full Screen View Mode

Summary

For the Full Screen view mode, we found a similar configuration as for the Hangout view mode and as shown in the like for like comparisons, there were also a high number of significant correlations between these two view modes. However, for the Full Screen view mode, the clustering is slightly more fragmented.



Detailed analysis

Table 1.3.2.3.3.3.: Full Screen View Mode significant correlations

| | VC | Track | Across | See | Lifelike | Close | Face | SimStarts | Silence | Lively | Contribute |
|--------------|----------|----------|----------|----------|----------|----------|----------|-----------|----------|----------|------------|
| VC | | | | | | | | | | | |
| Track | | | | | | | | | | | |
| Across | .020 | .000 | | | | | | | | | |
| See | | .080 | .003 | | | | | | | | |
| Lifelike | | .093 | .042 | .001 | | | | | | | |
| Close | | .000 | .000 | .001 | .000 | | | | | | |
| Face | | | .014 | .014 | .000 | .000 | | | | | |
| SimStarts | | .025 | | | | | | | | | |
| Silence | .059 | | | | | | | | .052 | | |
| Lively | | | | | .000 | .002 | .000 | .002 | | | |
| Contribute | | .000 | .000 | .009 | .010 | .000 | .006 | .006 | | .049 | |
| Total | 2 | 6 | 7 | 6 | 7 | 7 | 6 | 4 | 2 | 5 | 8 |
| p<.01 | | 3 | 4 | 4 | 4 | 7 | 4 | 2 | | 4 | 6 |
| p<.05 | 1 | 1 | 3 | 1 | 2 | | 2 | 1 | | 1 | 2 |
| p<.10 | 1 | 2 | | 1 | 1 | | | 1 | 2 | | |
| Size | 1.5 | 8 | 11 | 9.5 | 10.5 | 14 | 10 | 5.5 | 1 | 9 | 14 |

There were again many highly significant correlations, a total of 36.

1.4 Discussion

Vconnect aims to support ad hoc group video communication and closely map onto social network activities. This is different from deciding on a time to meet up via video conferencing to discuss specific topics. The main aim of this experiment was to decide on a user interface that could be beneficial when conducting the SAPO Campus first trials, with an option to switch lay-outs depending on the context of the mediated conversation.

We compared three different ways of displaying six participants in a highly lively discussion.

1. Tile View Mode, where six participants were displayed in two rows of three tiles.
2. A view mode similar to Google+ Hangout, driven by voice activity in such a way that the person who spoke was displayed in a larger window and the five non-speakers resided in five smaller windows (tiles) at the bottom of the screen.
3. Full Screen, where the active speaker would be displayed full screen.

It was thought that the full screen view mode would show facial expressions best and as such provide superior individual telepresence, although group cohesion might suffer. The tiled view mode would at all times show the group as a whole but it would be more difficult to see facial expressions of individuals. The Hangout style interface was thought to be a good intermediate solution as it would show facial expressions well and at the same time it would allow monitoring of the group as a whole.

In addition we wished to explore a slightly wider context and in interviews and questionnaires we probed about social network use, their current uses of video conferencing software and invited the participants to express their wishes and suggestions about interface design for social media inspired video conferencing.

Interviews

The feedback on the different view modes was at first glance in favour of both the Tiled and the Hangout view modes with very few positive comments about the Full Screen view mode.

The Tiled View Mode was well received in particular by the female participants. It supported awareness of the whole group at all times, and a lively group discussion with rapid fire turn taking as well as with more than one person talking.

The equal size of the tiles made participants feel that they were equal and part of the group which made it easier to contribute. They commented that the Tiled view mode felt more like an actual group discussions, mapping on well onto social networking aspects. The sense of group belonging was enhanced by being able to see oneself.

Paradoxically, in spite of the fast turn taking, the Tiled View Mode, where you could see everyone at all times also caused more awkward silences than the other two View Modes, as participants wanted to give other participants a chance to contribute. However, participants expressed that the Tiles should be bigger.

The Hangout View mode also received favourable comments. In particular it showed more facial details of a speaker whilst still being able to monitor the rest of the group.

An unexpected benefit was that the Hangout view mode seemed to help in those situations where you didn't know (many) others, because you could see better who was talking and you did not have to rely on recognising someone's voice.

Participants commented that the Hangout view mode might suit business meetings with slower turn taking better than more social network type of conversations.

However the comments highlighted that there are six problems that need to be remedied, under the following headings:

1. Tiles too small
2. Voice detection Vs. noise detection
3. Orchestration lagging
4. Animation too disruptive
5. Can't see your self
6. Location inconsistency

The small tiles at the bottom of the screen are too small and attenuate group awareness and cohesion.

Orchestration is driven by voice detection but on occasion Orchestration is determined by noises, e.g. heavy breathing. Orchestration is not always able to keep up with the rapid pace of conversational turns. In the Hangout view mode taking a conversational turn was emphasised by an animation where a speaker was relocated from its thumbnail at the bottom of the screen to the main window that is allocated to a speaker in a way that resembled a genie escaping from a bottle. With the lively conversation this popping up and out became disruptive. The inability to see one's self made people uncertain whether they would be seen and how they came across. In addition participants in the thumbnails did not keep their position in the tiles at the bottom which was disruptive.

A few male participants liked the Full Screen view mode, mostly because of the clarity of facial expressions. However not having a view on the rest of the group made it very difficult to keep track of the conversation. In addition the lag of the Orchestration engine was noticed much more in the Full Screen view mode, even though someone put its accuracy at 70%. It was difficult to monitor the whole group and



participants were not sure whether other more quiet participants were still there. In a way out of sight also meant out of mind. Participants did not feel part of the group anymore.

All the same, participants could think of scenarios where a Full Screen view mode would be appropriate: For small numbers of participants, for mobile phones and having multiple screens.

The View Modes experiment seemed to stimulate creative juices, some of them being quite solution oriented. There were two strands as to how to improve the Tiled view mode. The first one was to simplify the interface even further to enhance the group feeling by removing the boundaries of the tiles. The second was to highlight the tiles of those who were speaking.

Vconnect intends to facilitate easy joining and leaving of group video chat. The participants had some interesting suggestions. There is an awareness that interfaces need to be simple.

Currently our participants are not in the habit of carrying out ad hoc group video chats. There are certain obstacles in organising these. But on the few occasions they have a group video chat it is, in the words of one participant, heaps of fun. They can see the need to get in touch with people you don't see often, but can't imagine a need to get in touch again with people you see all day, unless it is about homework. The latter fits well with a SAPO-Campus scenario, a social network within an educational environment. As such these trials formed an excellent preparation for the up and coming field trials at SAPO-Campus (Portugal Telecom).

Statistical Analysis

The statistical analysis showed a great many highly significant differences and similarities and led to some clear and well supported interpretations. As such we can have some confidence in being able to generalise to a wider population of social network users and what opportunities there are for ad hoc group video conferencing.

Overall the participants preferred the Tiled view mode; importantly well over 80% of the females preferred the Tiled view mode. There is a considerable group though that prefers the Hangout style view mode, mostly consisting of males. Participants who are experienced in video conferencing tend to have a preference for Hangout. We can consider the Hangout view mode as a close second. All the females and over 83% of males liked the Full Screen view mode least. Most of the participants were intensive social media users.

The Tiled view mode proved to be significantly better than the Hangout view mode for keeping track of the conversation. The Tiled view mode completely outperformed the Full Screen view mode as it was significantly easier to keep track of the conversation, participants seemed more lifelike, it was easier to contribute to the conversation, it was easier to see who was talking, participants felt they came across better and felt closer to the others. In other words in the Tiled view mode they felt much more part of the group and this facilitated being part of the conversation.

Although the level of significance was slightly lower, the Hangout view mode also outperformed the Full Screen view mode, as participants also seemed more lifelike, facial expressions were clearer, they could see better who was talking, they felt closer to the others, it was easier to keep track of the conversation and they felt they came across better. It is not as if the full screen mode did not accommodate seeing facial expressions. In the interviews there are references to appreciating being able to see facial expressions in the Full Screen view mode. However, in a group discussion it is so important to see the others at all times. In the Tiled view mode and to a lesser extent in the Hangout view mode they could still see all the others.

In the Hangout view mode, previous experience with videoconferencing helped participants feel closer to each other, made it easier to contribute to the discussion and they were better able to keep track of

the conversation. As the latter in particular was so much easier in the Tiled view mode, it seems that with a bit of practice certain barriers can be overcome.

63% of the participants knew all or most of their fellow conversational partners. Those who did not know many or none at all were affected in the Hangout view mode in a surprising manner. Those who knew few or none of the other participants were better able to keep track of the conversation, see other participants, they felt closer to others, experienced fewer simultaneous starts and they found it easier to contribute. It is not obvious why this happened, but the Hangout view mode seems to accommodate people who know fewer people well.

Cluster analysis revealed that experience with video conferencing did not have any bearing on user experience in the Tiled View mode but the intensity of using social networks did. Those who reported to use social media more intensively also found it easier to contribute to the conversation, found the other participants more lifelike, could see better who was talking and saw facial expressions better. Almost all the variables showed a high number of extremely significant inter-correlations. We were able to plot all the questions in a two dimensional space where the X-axis reflected the level of conversational flow and the Y-axis ran between “individual telepresence” and “group telepresence”.

For the Hangout and Full Screen view modes, in contrast to the Tiled view mode, experience with video conferencing correlated significantly with eight of the Hangout view mode ratings but the intensity of using social networks did not yield any significant correlations. This suggests that the experience in the Hangout view mode is quite different from using the Tiled view mode. The Tiled view mode seems to map on to the (intensity of the) use of social networks and the Hangout view mode on the use of video conferencing.

It seems tempting to paraphrase the findings as that the Tiled view mode emphasises socialising interactively with people, possibly an explanation as to why in particular females prefer this view mode, whereas the Hangout view mode emphasises talking about “things” [Whitaker, 1995] where previous experience with video conferencing is of importance.

To conclude

The experiment simulated group social interaction resulting in fast turn taking. The interviews and questionnaire clearly showed that the Tiled view mode was better (and preferred) at supporting highly interactive, lively, social interaction. The cluster analysis also showed a very well defined (strongly inter-correlated) space for the Tiled view mode, with the Y-axis running between individual and group telepresence and the X-axis signifying levels of conversational flow, or in simple terms between slow and fast turn taking (figure 1.4.1.).

The Tiled view mode then was able to support fast turn taking and supported both group and individual telepresence. In addition, the Tiled lay-out would allow easy (ad-hoc) joining and leaving of a group if it were part of a social network page.

The Hangout view mode was regularly mentioned in the interviews as a more *formal* space better suited to slower turn taking, emphasising individual telepresence, although there was still opportunity to monitor the group as a whole.

The Full Screen view mode did not accommodate group interaction at all, although participants mentioned that they liked the Full Screen view mode for seeing facial expressions of an individual. This led them to hypothesise that for small screens and for a small group of conversational partners the Full Screen view mode might work well.

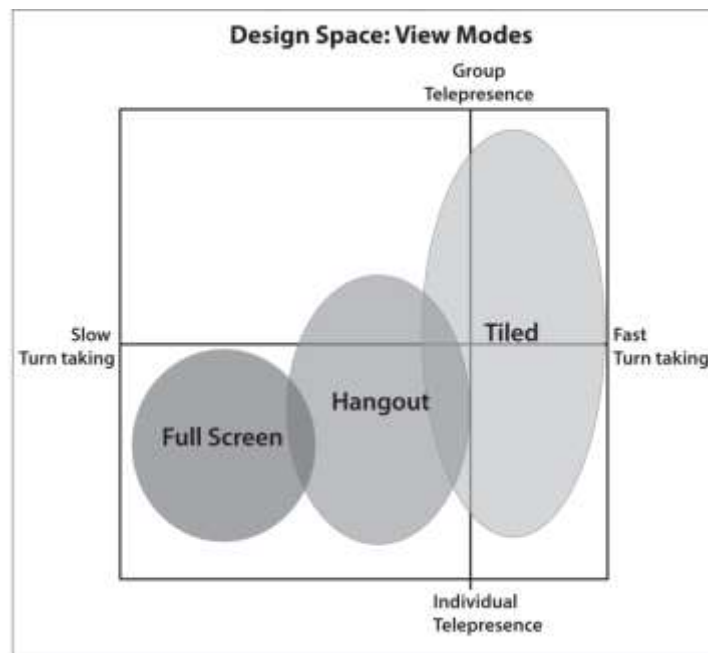


Figure 1.4.1.: Design Space View Modes

In a similar experiment [reported in chapter 2] we compared Full Screen Orchestration with a Static Group view using large size televisions in a living room scenario (with two persons on a sofa) between three remote locations. We found very few significant differences between the two view modes. There was some indication that it was harder to keep track of the conversation in the Full Screen view mode but conversely using (Full Screen) Orchestration to take full advantage of the pan-zoom-tilt cameras the effect of close-ups and details of facial expressions delivered some dramatic effects enhancing telepresence. Somehow the Full Screen view mode does not translate from the large TV living room scenario to the smaller (computer-)screen when there is only one person per location.

That is not to say that there are no good opportunities for Full Screen view modes in the wider landscape. We already mentioned the example of a small screen application; mobile phone video conferencing with a small number of participants. On the other end of the screen size continuum, it may well be that there is a good opportunity to raise the dramatic temperature in a performance scenario using large Full Screen orchestration. More generally, in any situation where the nature of the dialogue is one of slow turn taking, e.g. seminars, Full Screen Orchestration might be desirable.

Vconnect aims to support ad-hoc group videoconferencing via computer screens (mostly involving one person per location) and the current experiment highlights that in this scenario a Full Screen view mode is not optimal. In addition Vconnect aims to integrate with and even map on onto social networking, rather than providing yet another computer based video conferencing system. In this respect it is advisable to listen to the females in our study, as it is still true that by and large women talk about people and men talk about things.

The Tiled view mode with its consistency of location for each participant seemed to map very well onto social networking and with its emphasis on equality and being able to monitor all group members easily in such a consistent manner, has the potential to fulfil Vconnects aim of supporting ad hoc group video conferencing. However, the Tiled view mode, as it stands, offers little in being a platform or showcase for Vconnect technologies.

The Hangout style detracts from group awareness by changing the location of participants continuously in the small tiles below the main window. This is an issue that can easily be addressed. It seems beneficial to incorporate the insights of the last few decades of human factors research into the importance of the non-speaker in group discussions, the importance of consistency in lay-out, i.e. each participant should remain in the same location on screen during a session. In the interviews participants suggested that the Hangout view mode would benefit from a smaller main window (for the person who speaks) and bigger tiles for the non-speakers. Again this can easily be rectified and the Vconnect technical team has already made changes in this regard.

All things considered there are only marginal differences between the Tiled and Hangout view modes and the poorly functioning parts of the Hangout view mode can easily be repaired. In addition the Hangout view mode offers better opportunities to incorporate sharing digital artefacts, such as sharing a video, a picture and so on.

Recommendation

It is recommended then to use the Hangout style view mode for further technology development but make changes to the Hangout style view mode in line with what this experiment uncovered, i.e. reduce the size of the main window, increase the size of the tiles, introduce screen-location consistency, and remove the current animation feature (going from tile to main window and back). Most of these changes have been made now, a good metric of how integrated the user research is into the technology development.

The example design in figure 1.4.2. has grouped six participants together (without borders) to enhance a feeling of group closeness, where the size of the tiles are big enough to afford group and individual telepresence. The bigger box can be used for a shared focus, e.g. media sharing or more to the point an area where orchestration can be used, e.g. on occasions where one or more persons present, or a more formal video conference occasion requiring slow turn taking.



Figure 1.4.2. Example Design



2 View Modes in the Living Room

Goldsmiths: Emmanouil Falelakis, Michael Frantzis, Marian Ursu, Vilmos Zsombori
Falmouth University: Erik Geelhoed

2.1 Introduction

Most of the introduction we used for Chapter 1 (Viewmode Experiment) also applies to this section. Here we report back on a laboratory experiment in which three “living-rooms” were connected and where we asked participants (two in each room) to engage in a highly lively and social exercise under two conditions.

The first condition involved using one static camera per room and the participants in each room could see the other participants in the remote rooms as a group. This was thought to afford good group awareness, including being able to identify where vocal backchannels came from but would fall short on feedback through facial expressions.

The second, Orchestrated condition, made use of three cameras, of which one was a Pan-Zoom-Tilt (PZT) camera and Vconnect voice driven technology to display full screen the (zoomed-in) active speaker interspersed with zoomed-out group shots. Here facial expressions of the person who talked would be clearly visible and was thought to enhance telepresence of the speaker, but group awareness might suffer as there would be no continuity in showing the group as a whole.

2.2 Method

2.2.1 Participants

Twenty four participant, 18 female (mean age 22.89, SD 3.86), 6 male, (mean age 22.83, SD 3.25) were recruited from Goldsmiths' student population. There were four experimental sessions and in each session six students took part.

2.2.2 Studios, displays, cameras

Three rooms at Goldsmiths all three measuring about 12 m² were connected via the video conferencing system. Although for practical reasons the rooms were in close proximity, there was no natural sound leakage from one room to another. The rooms featured comfortable seating in such an arrangement that directional audio could be determined accurately.

Each room featured one 50" TV display and participants sat at a distance of about 2m from the display. One Sony EVI-HD1Pan Zoom & Tilt camera was positioned centrally just below a display and two Panasonic AG-AC160 (fixed) cameras were placed on tripods on either side of a display close to the bottom of a display.

The Return Transmission Time (RTT), the standard measure of roundtrip delay, was found to be constant and of the same value between all three rooms, and measured 650 milliseconds.

2.2.3 Vconnect System

In each room the direction of the person speaking was determined and this cue was then processed via a rules based system which resulted in automatically editing / orchestrating the video output for each room, which in turn resulted in different composition for each room [see section 2.2.6.2). Participants were not able to see

themselves, e.g. via Picture in Picture (PIP) nor were they able to derive feedback as to how (or if) they were shown on the screens of the other two rooms. Directional audio was determined using an array of four condenser microphones.

2.2.4 Experimental conditions, order, procedure

There were two experimental conditions:

1. Orchestrated using the three camera Vconnect system (Figure 2.2.4.2.)
2. Single static camera (Figure 2.2.4.1.)

Each experimental session involved two participants per room (six per session). All participants took part in both conditions in one of two orders:

1. Order 1: Static condition followed by orchestrated/Vconnect condition
2. Order 2: Orchestrated condition followed by the static condition



Figure 2.2.4.1: Static Condition



Figure 2.2.4.2: Orchestrated Condition

In the static condition a screen showed one remote room at the top and the other at the bottom of the screen in a geometrically consistent configuration, i.e. if the participants in room 1 see room 2 at the top and room 3 at the bottom then the participants in room 3 should see room 1 at the top and room 2 at the bottom. The participants in room 2 see room 3 at the top and room 1 at the bottom.

2.2.5 Task

For one condition, participants were asked to generate a list of seven items between them pertaining to "what constitutes an ideal holiday" using a maximum of 5 minutes and then using another maximum of 5 minutes to prioritise and agree upon a group list of these seven items. A (deliberately) small piece of paper and a pencil was provided to each participant.

For the other condition the task was to generate and prioritise seven items pertaining to an "ideal home". A new small piece of paper was provided.

The topics for the task are relatively neutral and people of different backgrounds are equally able to carry the task out. In addition there is (practically) no learning curve, as such participants perform equally well in the first as in the second session.

The task is highly repeatable across a number of synchronous and a-synchronous experimental paradigms, i.e. the experimental validity is high. In addition, the tasks elicits highly interactive behaviour around trivial topics in a way that mimics (some) communication through social media and thus we might argue that there is a reasonable ecological validity.

2.2.6 Measures

1. Short unstructured group interviews after the experiment (i.e. after both conditions), asking about likes, dislikes and wishes of the technologies used.
2. Short questionnaires in graphic rating scale format (Stone et al 1974) after each condition. The graphic rating scales measured 112 mm.
3. Automated logging

2.2.6.1 Questionnaires

For both conditions the participants answered 16 identical rating scale questions about aspects of telepresence. Below is an example of a rating scale question:

- How close did you feel to people in the other rooms?

Not at all

Very

|_____|

Participants were asked to make a mark on the scale between (and including) the two extremes.

The questions asked about well-established aspects of telepresence, adverse effects on conversational parameters and group interaction:

1. How much was looking at the people in the other rooms like looking through a window?
2. How much was looking at the people in the other rooms like watching a TV panel discussion?
3. How lifelike were the people in the other rooms?
4. How close did you feel to people in the other rooms?
5. How much did you notice the facial expressions of people in the other rooms?
6. How much did you feel you had eye contact with people in the other rooms?
7. How much did it distract from the group conversation when you spoke to the person sitting next to you in your room?
8. How well could you see the persons in the other rooms?

9. How much did you notice a delay in the communication, i.e. it looked like the people in the other rooms heard you a little while after you spoke?
10. How disruptive did you find the delay?
11. How often did it happen that someone in your room and someone in one of the other rooms started talking at the same time?
12. How often were there awkward silences between the three rooms?
13. How lively were the discussions between all of you?
14. How similar to a face-to-face meeting was this session?
15. How easy was it to keep track of the discussion?
16. How well did you feel you came across to the other rooms

The questionnaire data were analysed using SPSS (Statistical Package for the Social Sciences, IBM) identifying statistical descriptions, analysis of variance was used to explore differences, correlations and Multi Dimensional Scaling were used to analyse similarities.

2.2.6.2 Automatic logging

The operation of the orchestration engine is essentially based on the continuous realisation of two discrete activities¹, as illustrated in Figure 2.2.6.2.1:

- Analysis of low-level cues from different sources and/or modalities in order to *understand* important events that occur during the interaction and
- Use of appropriate, pragmatic and aesthetic, principles in order to *react* by choosing the most appropriate shot/viewpoint to display at each point or, in other words, to represent the interaction on a screen.

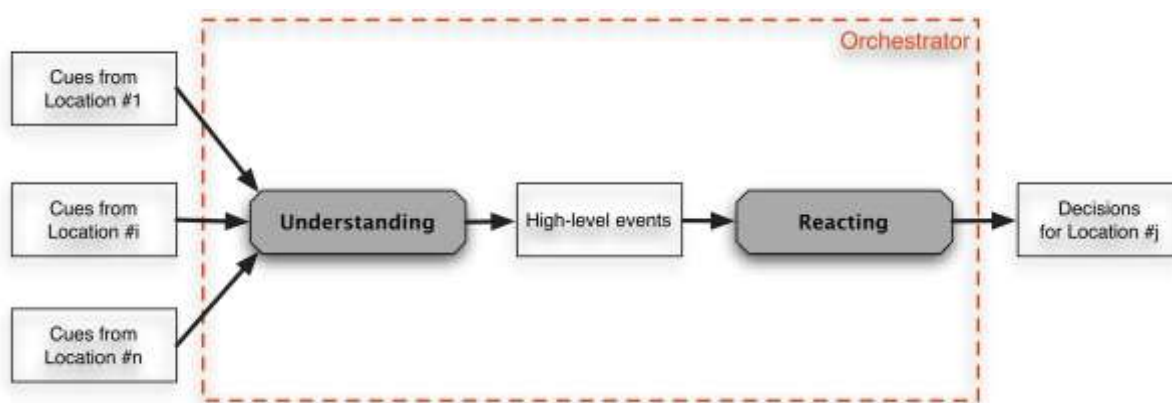


Figure 2.2.6.2.1 The Orchestration pipeline

In the context of the current experiment, low-level data comprised voice activity cues for each participant in each room. These were fused by the Understanding module and interpreted into the following high-level events:

¹ For details regarding the orchestration reasoning framework and its implementation, the reader is referred to the Vconnect deliverables *D3.6 - Interim conceptual frameworks for expressing orchestration knowledge* and *D3.3 – Communication reasoning engine first release*, respectively.

- **Turn shift:**
A turn shift occurs when there is significant (i.e. lasting more than 400ms) voice activity from a person who does not currently have the turn, excluding the case of a cross talk. The turn is maintained, even if there are some short (i.e. lasting no more than 800ms) gaps in the speech of a participant.
- **Cross talk:**
A Cross-talk occurs when a person starts speaking at a point when another person has the conversational turn.
- **Pattern of short-turn taking:**
A pattern of short-turn taking occurs when there are 3 or more turn shifts over a period of 6 seconds

Based on these events, and using a set of adequate rules, the Reacting module continuously produced the following two types of events:

- **Shot proposal:**
A shot proposal is made by the Reacting module whenever the predicates of a mixing rule are satisfied. As its name suggests, a shot proposal is not necessarily accepted and cause an actual cut. This is because it competes for the screen with other shot proposals produced concurrently.
- **Cut**
A cut is necessarily a change in the visual representation or, in other words, a shot proposal that has been accepted.

During the experiment, all the events presented above have been recorded, as illustrated in Figure 2.2.6.2.1.

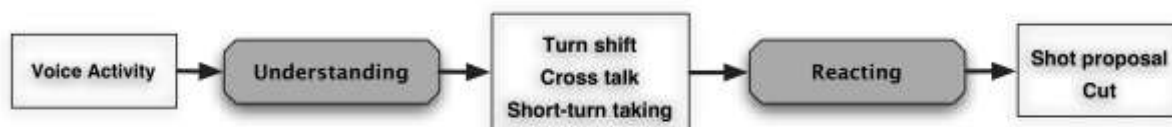


Figure 2.2.6.2.1 Automatically logged events in the orchestration pipeline

Based on the logged events, for each experimental session, a number of metrics have been calculated and the results are presented in this document, namely:

1. Average number of turn shifts per condition and per session
2. Total number of turn shifts per room
3. Average number of short turn takings per condition and per session
4. Total number of short turn takings per room
5. Average number of cross talks per condition and per session
6. Total number of cross talks per room
7. Average turn duration per condition and per session
8. Average turn duration per room
9. Average number of voice activities per condition and per session
10. Total number of voice activities per room
11. Average number of shot proposals per condition and per session
12. Total number of shot proposals per room
13. Average number of cuts per condition and per session
14. Total number of cuts per room
15. Average shot duration per condition and per session
16. Average total duration of shots per room

2.3 Results

2.3.1 Interviews

Group interviews were carried out with all the participants after they had completed all the experimental conditions. The following describes some of the themes that emerged from these discussions with potential feedback into future iterations of an orchestration logic.

Orchestration with different shot sizes segments the personal space as well as the visual space

A noticeable effect of using close up shots on individual people was that it seemed to create an effect of segmenting the space on a personal level i.e. gave participants the impression that they were in separate spaces, as well as on a visual level. So whilst it was possible for individuals to feel closer to an individual with close up shots – I could “... *feel more closer to them... more personal*” - it could make it feel like a separate conversations – *‘it meant a series of one on ones’*. One participant stated – *“I always felt like I could always turn around to my partner in the room and have like a private conversation”*. Another said - *“It’s turning a social setting into these one on one things - I’m not comfortable with that. I want to see other people’s reactions to ideas”*.

This points to successful orchestration needing to take into account how the screen layout can affect the segmentation of the personal space. For example these comments would suggest using a split screen of multiple shots might have a more unifying effect on a group rather space than full screen cutting of close up shots which creates a more private one-one-effect. Being able to choose screen layouts along such coordinates could possibly make orchestration more effective.

The effect of this choosing of close up shots did seem unpredictable with some participants not finding the effect more intimate. Some found the cutting of shots itself quite distracting. Others even found the errors in orchestration (when often it wasn’t showing who was talking) helped in creating a sense of equality by showing more of the non-speakers. Others again experienced an effect of not needing to look at the screen as much – a liberating effect.

The problems of representing multiple people in multiple screens – where do you look?

However, it is not necessarily the case that a split screen layout was the preferred option for creating a group experience. For example, some interviewees expressed the sentiment that it was hard to know where to look when a number of people were talking at the same time on split screen. All the participants appeared smaller and it was less obvious.

‘When am I on screen?’ and self view?

“In normal conversation, you ‘just speak’ you don’t wonder if the camera is on you.”

Whilst some of the idea behind orchestration is to create an immersive experience the nature of the context, communicating across TV screens and edited vision mixes, inevitably introduces a different dynamic to interactions when compared to talking together in a group. In a group it seems like a natural thing to start a conversation and expect to be seen and heard. This is no longer the case the context of an orchestrated video conference. Some participants adapted to it, enjoyed it, no longer noticed it. Some felt the use of different shot sizes selecting individual people made it more like a regular group experience by reflecting the fact that we look at people in turn, others found it had the opposite effect.

A particular negative effect was that with changing shots, some participants felt the loss of knowing when they appeared on screen. If they could see editing, but they didn’t know what was happening on other people’s screens, then they found it difficult to know if they would be able to attract attention in the conversation. They were feeling excluded.

Related to this were attitudes towards self-view. Some participants felt that a self view would help with this problem however others felt no need for it and could absorb themselves more easily in the experience. Maybe this needs to be an option for users?



Intimacy and formality

Related to the above is the fact that full screen orchestration seemed to provide a more intimate experience compared to the more formal one offered by a static split screen composition. The static screens were better for getting things done it was felt, but the orchestrated close ups made it more fun and intimate. This builds on the idea of the use of close up shots creating a sense of a series of one on one's type conversation i.e. a series of more intimate interactions. With orchestration – “it was quite intimate seeing other people's faces” but for the static shot – “it was like trying to get stuff done”.

One participant summed up orchestration as being more ‘real’ than Skype “because you could see people closer and mainly you could see the facial expressions”.

This would imply orchestration, in conjunction with control of the screen layout, could choose different screen compositions and different orchestrating styles according to the context of the interaction e.g. formal class room or simple supervision, or formal meeting versus friends getting together.

The importance of the audio channel and how well participants know each other

When communicating with an orchestration system that seemed to be functioning less well, one participant commented - “I try to match what I see with what I hear...”. The current implementation of the orchestrator relied upon voice activity primarily as a cue, with these audio cues being further interpreted into conversational turns. However, rather than merely using the audio channel as a source to extract cues from, orchestration maybe needs to include the audio as an integral part of the experience and of the systems understanding of the experience.

For example, if people already know each other then it is easier for them to notice recognize each other from the audio. This would mean that it is less crucial for the correct speaker to be shown on screen at all times when participants in a video conference know each other compared to when they do not. This additional information contained in the audio channel may be able to mitigate the segmentation of the space resulting from the use of close ups described earlier if the people know each other and allow for the benefit of the intimacy which come with close up shots without compromising too much on the group experience. How well participants know each other, then needs to become an input into the orchestrator knowledge. This input, i.e. knowledge of how well people know each other, would be an input that would need to come from a source outside of the immediate space of the interaction and using this input would mean orchestration would be varying the experience according to the communication context.

Changing the screen layout according to context

This opens the door to varying orchestration, shot choices and screen layout, according to how well the people involved know each other. For example, if 1 person out of 6 happens to be on one Vclient on his own then he might receive a split screen showing 5 small close ups of the other participants to allow him/her to orientate themselves better in the communication whilst the other would receive full screen orchestrated feeds.

Also, when a person joining a new conference might be unknown to the other participants then a period of close ups of the other participants might help the process of introduction before settling down on a tiled layout.

Dealing with cross talk

Another limitation of full screen close-up orchestration is that, in a situation of cross talk it is not possible to show both people talking and reactions to talking when more than one individual is talking at the same time. This does tend to bother respondents, though not universally, with some viewing the orchestration condition as more relaxed.

However, this in itself could possibly provide an additional cue into screen layout choices. For example, if there were a continued period of cross talk then the orchestrator could maybe choose a layout that best represented the two relevant people to the other.

Keeping up with the speed of interactions

Similar to the problems experienced with cross talk was, with the steadily increasing number of participants, the probability of a rapidity of communication turns increased. This posed a similar problem to cross talk in that following audio activity communication turns became impossible. As one participant put it - “The image cannot keep pace with our conversation”

The existing orchestration logic had built into it a limiter which prevented an edit being made before an existing shot has been on screen for eg. 2 secs. This breaks down when conversation turns are shorter than 2 secs.

A possible input into the orchestrator could be the average length of the turn taking place in the conversation. This could then provide a weighting of whether to orient towards a tiled layout or full screen layout i.e. if turn taking is too rapid then use a more static tiled or split screen layout.

2.3.2 Automated logging

2.3.2.1. Turn-shifts

A turn shift occurs when there is significant (i.e. lasting more than 400ms) voice activity from a person who does not currently have the turn, excluding the case of a cross talk. The turn is maintained, even if there are some short (i.e. lasting no more than 800ms) gaps in the speech of a participant.

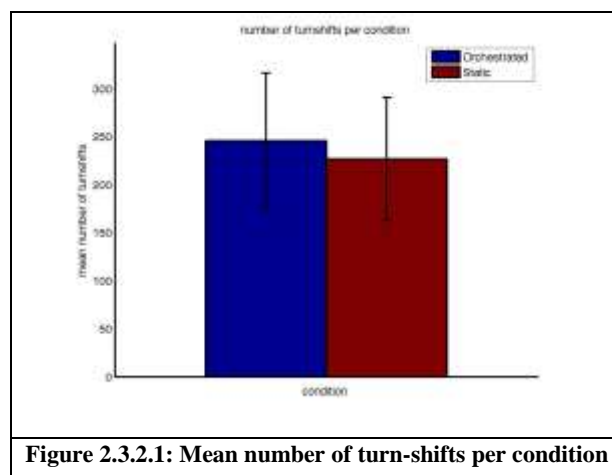


Figure 2.3.2.1: Mean number of turn-shifts per condition

Table 2.3.2.1: Turn-shifts across sessions and conditions

| Sessions | Orchestrated | Static | χ^2 (df 1) | P |
|--------------|--------------|--------|-----------------|-------|
| 1 | 143 | 133 | 0.36 | n.s. |
| 2 | 282 | 264 | 0.59 | n.s. |
| 3 | 258 | 239 | 0.73 | n.s. |
| 4 | 300 | 271 | 1.47 | n.s. |
| Total | 983 | 907 | 3.06 | < .10 |
| Mean | 245.75 | 226.75 | | |

Table 2.3.2.1 shows for the four sessions (1 – 4) the number of conversational turns in the orchestrated and static conditions. Comparing the frequencies using Chi Squared, we found that there were no significant differences per session, although in the orchestrated condition there seem to be more conversational turns than the static one. This pattern appears to be consistent across the sessions. This leads to the total of conversational turns across sessions in the orchestrated condition being greater than the static condition (figure 2.3.2.1.), although the level of significance is only 10%.

One interpretation would be that in the orchestrated condition participants are more animated and tend to take more turns. This could be the effect of continuous shot changing somehow influencing the pace of the conversation.

2.3.2.2. Short-turns

A pattern of short-turn taking occurs when there are 3 or more turn shifts over a period of 6 seconds.

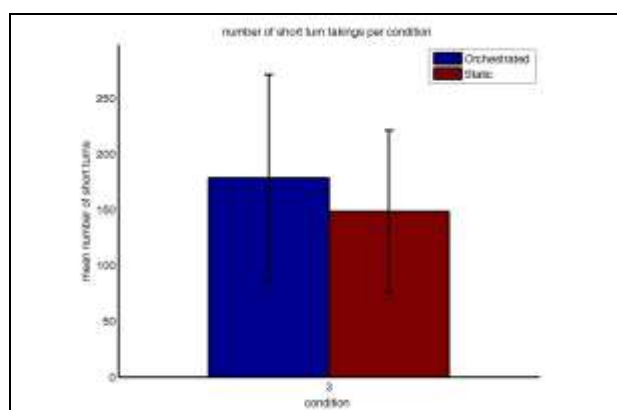
**Figure 2.3.2.2: Mean number of short-turns per condition**

Table 2.3.2.2: Short Turns across sessions and conditions

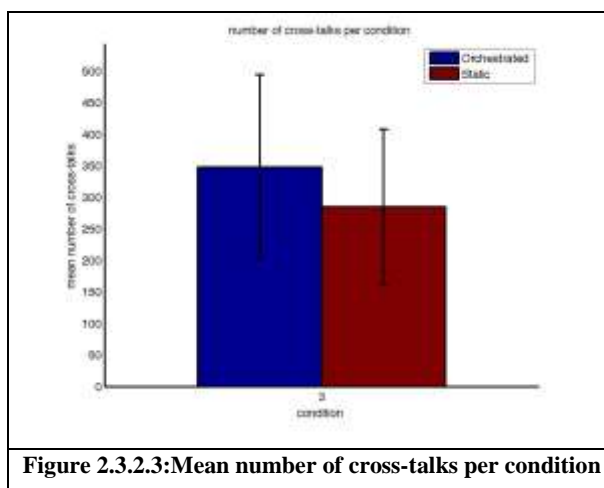
| Sessions | Orchestrated | Static | χ^2 (df 1) | P |
|--------------|--------------|--------|-----------------|------|
| 1 | 45 | 45 | 0 | ns |
| 2 | 234 | 188 | 5.01 | <.05 |
| 3 | 187 | 154 | 3.19 | <.10 |
| 4 | 248 | 208 | 3.51 | <.10 |
| Total | 714 | 595 | 10.82 | <.01 |
| Mean | 178.5 | 148.75 | | |

The number of short turns in the two conditions, with the exception of the first session where there numbers are equal, shows a consistent higher number of short turns in the orchestrated condition for session 2-4 (With session 2 being significant at 5%) and a highly significant ($p < .01$) for the totals across sessions. A high frequency of short turns is indicative of a lively (group) conversation.

2.3.2.3. Cross Talk

A Cross-talk occurs when a person starts speaking at a point when another person has the conversational turn.

With exception of the first session, even stronger are the differences for cross-talks, number of interruptions, highly significantly higher in the orchestrated condition. Although consistent with liveliness of the conversation, it could also signify an attempt by the group to keep track of the conversation.


Figure 2.3.2.3: Mean number of cross-talks per condition
Table 2.3.2.3: Cross-talks across sessions and conditions

| Sessions | Orchestrated | Static | χ^2 (df 1) | P |
|--------------|--------------|--------|-----------------|-------|
| 1 | 328 | 298 | 1.44 | n.s. |
| 2 | 548 | 442 | 11.35 | <.001 |
| 3 | 196 | 145 | 7.63 | <.01 |
| 4 | 320 | 254 | 7.59 | <.01 |
| Total | 1392 | 1139 | 25.29 | <.001 |
| Mean | 348 | 284.75 | | |

2.3.2.4. Turn Duration

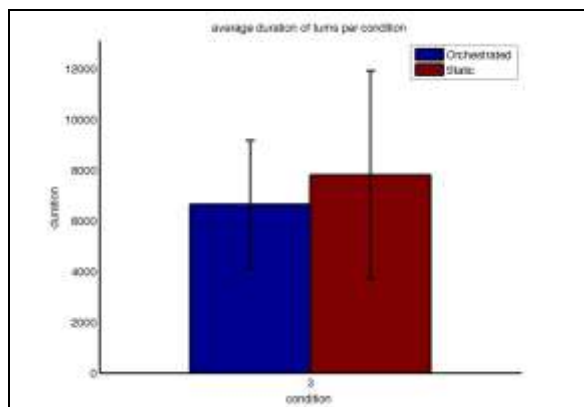


Figure 2.3.2.4: Mean duration of turns per condition

Table 2.3.2.4: Turn duration across sessions and conditions

| Sessions | Orchestrated | Static | F _(1,3) | p |
|--------------|--------------|---------|--------------------|------|
| 1 | 10354 | 13949 | | |
| 2 | 4962.2 | 6166.9 | | |
| 3 | 6129.5 | 5265.5 | | |
| 4 | 5162 | 5925.2 | | |
| Total | 26607.7 | 31306.6 | 1.626 | n.s. |
| Mean | 6651.925 | 7826.65 | | |

Turn duration is not expressed as a frequency. Treating the numbers as scale data, we carried out a repeated measures analysis (Orchestrated Vs Static). This did not result in a significant difference between the two conditions.

2.3.2.5. Voice Activity

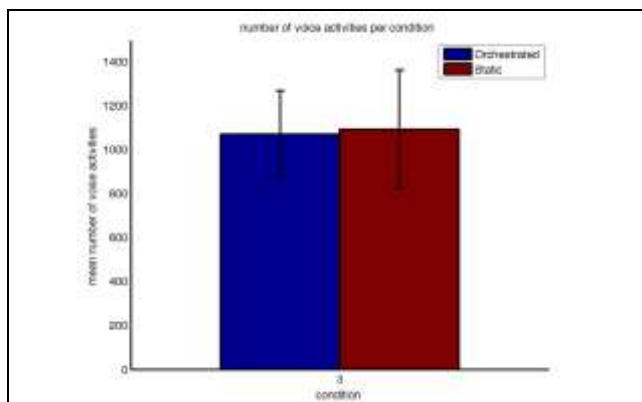


Figure 2.3.2.5: Mean number of voice activities per condition

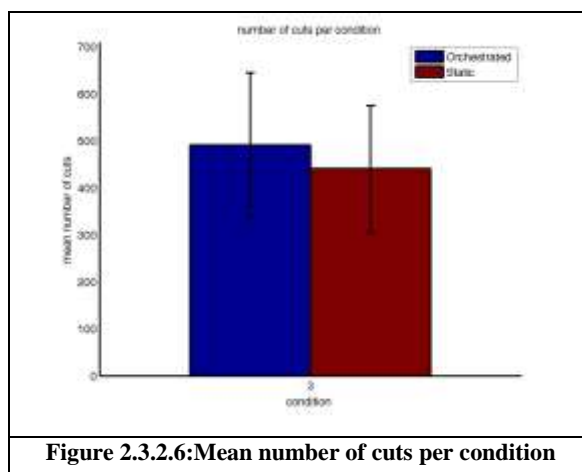
Table 2.3.2.5: Voice Activity across sessions and conditions

| Sessions | Orchestrated | Static | χ^2 (df 1) | P |
|--------------|--------------|--------|-----------------|--------|
| 1 | 989 | 1412 | 74.52 | <..001 |
| 2 | 1274 | 1203 | 2.04 | n.s. |
| 3 | 831 | 806 | 0.38 | n.s. |
| 4 | 1183 | 951 | 25.22 | <.001 |
| Total | 4277 | 4372 | 1.04 | n.s. |
| Mean | 1069.25 | 1093 | | |

There were no consistent patterns for the number of voice activities per se. In session 1 there are significantly more voice activities registered in the static condition, in session 4, this is the reverse, whilst sessions 2 and 3 did not differ significantly.

2.3.2.6. Cuts

A cut is necessarily a change in the visual representation or, in other words, a shot proposal that has been accepted.


Table 2.3.2.6: Cuts across sessions and conditions

| Sessions | Orchestrated | Static | χ^2 (df 1) | P |
|--------------|--------------|--------|-----------------|-------|
| 1 | 279 | 260 | 0.67 | n.s. |
| 2 | 601 | 555 | 1.83 | n.s. |
| 3 | 481 | 418 | 4.42 | <.05 |
| 4 | 606 | 531 | 4.95 | <.05 |
| Total | 1967 | 1764 | 11.05 | <.001 |
| Mean | 491.75 | 441 | | |

Overall there are significantly more cuts made in the orchestrated condition (and in the static condition there were no actual cuts of course), although there are no significant differences in the first two sessions.

2.3.2.7. Shot Proposals

A shot proposal is made by the Reacting module whenever the predicates of a mixing rule are satisfied. As its name suggests, a shot proposal is not necessarily accepted and cause an actual cut. This is because it competes for the screen with other shot proposals produced concurrently.

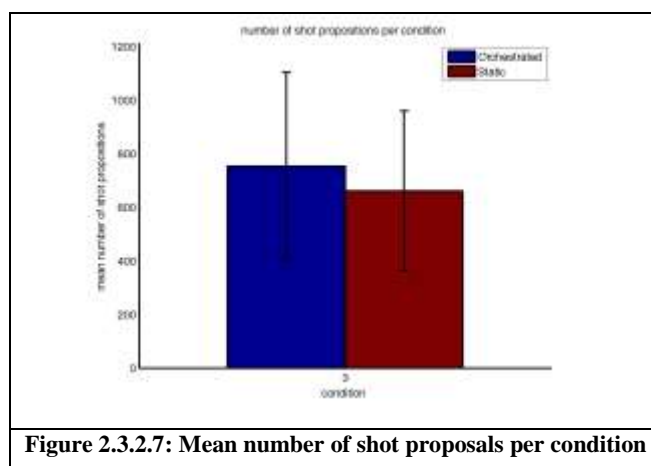


Figure 2.3.2.7: Mean number of shot proposals per condition

Table 2.3.2.7. Shot proposals across sessions and conditions

| Sessions | Orchestrated | Static | χ^2 (df 1) | p |
|--------------|--------------|--------|-----------------|-------|
| 1 | 286 | 266 | 0.73 | n.s. |
| 2 | 1098 | 967 | 8.31 | <.01 |
| 3 | 712 | 620 | 6.35 | <.02 |
| 4 | 919 | 792 | 9.43 | <.01 |
| Total | 3015 | 2645 | 24.19 | <.001 |
| Mean | 753.75 | 661.25 | | |

With the exception of session 1, the number of shot proposals is significantly higher in the orchestrated condition.

2.3.2.8. Duration of shots

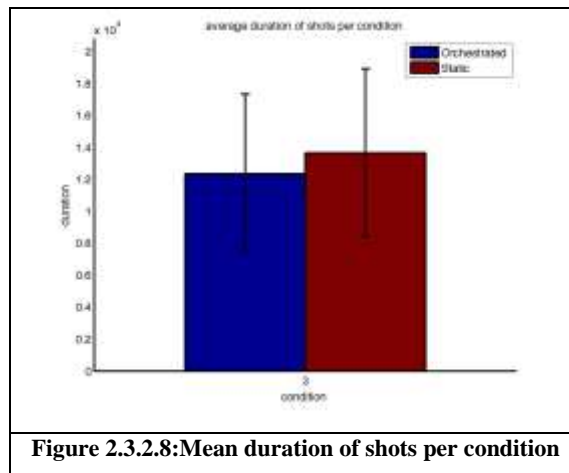


Table 2.3.2.8. Shot duration across sessions and conditions

| Shot duration | | | | |
|---------------|--------------|----------|--------------------|------|
| Sessions | Orchestrated | Static | F _(1,3) | p |
| 1 | 19589 | 21118 | | |
| 2 | 8995.3 | 9630.7 | | |
| 3 | 11699 | 13586 | | |
| 4 | 9085.6 | 10329 | 25.041 | .015 |
| Total | 49368.9 | 54663.7 | | |
| Mean | 12342.23 | 13665.93 | | |

Again treating the data as scale data, carrying out a repeated measures ANOVA, the duration of the orchestrated shots is significantly lower.

2.3.3 Questionnaires

The graphic rating scales measured 112 mm. To present the results on a scale of 1 – 100, we multiplied the raw data by $100/112 = 0.892857$.

2.3.3.1 Descriptive Statistics

2.3.3.1.1 Descriptive Statistics Static Condition

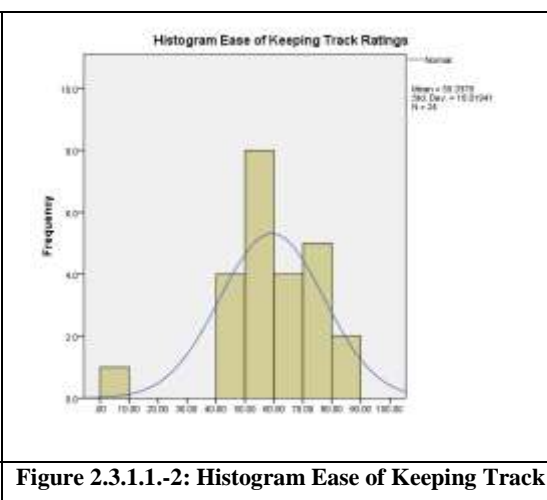
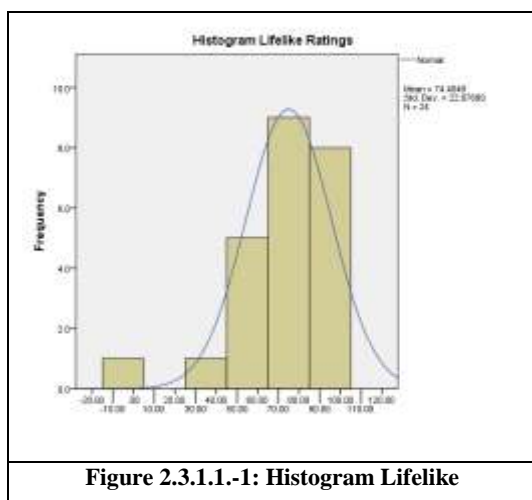
Table 2.3.1.1-1. shows descriptive statistics for the static condition ordered from high to low means. The full list of questions can be found under section 2.2.6.1. In the tables and graphs below abbreviations are used, which hopefully are self-explanatory.

Table 2.3.1.1-1: Static Condition: Descriptive Statistics

| | Minimum | Maximum | Mean | Std. Error | Std. Deviation | Skewness | Kurtosis |
|----------------------------|---------|---------|----------------|------------|-----------------|----------|--------------|
| Lifelike | .00 | 100.00 | 74.4048 | 4.66973 | 22.87689 | -1.518 | 3.628 |
| Simultaneous starts | 27.68 | 100.00 | 72.8795 | 4.21736 | 20.66076 | -.940 | .236 |
| Lively | 35.71 | 100.00 | 70.8333 | 3.14695 | 15.41685 | -.435 | -.050 |
| See others | 33.04 | 92.86 | 67.1875 | 3.70841 | 18.16743 | -.119 | -1.222 |
| Facial expressions | 23.21 | 91.07 | 60.1935 | 4.44870 | 21.79411 | -.267 | -1.234 |
| Keeping track | .00 | 87.50 | 59.3378 | 3.67820 | 18.01941 | -1.324 | 4.038 |
| Come across | 11.61 | 91.07 | 55.3571 | 4.62789 | 22.67194 | -.291 | -.785 |
| Close to others | 8.04 | 97.32 | 54.7360 | 5.70754 | 27.37241 | -.170 | -1.107 |
| Like tv-panel | 1.79 | 83.93 | 54.0551 | 5.21452 | 25.54581 | -.765 | -.831 |
| Side conversation | 5.36 | 83.93 | 48.0283 | 4.90335 | 24.02141 | .018 | -1.231 |
| Like face to face | 5.36 | 84.82 | 45.6101 | 4.94089 | 24.20531 | -.043 | -1.016 |
| Like a window | .00 | 83.93 | 42.3913 | 5.27710 | 25.30809 | -.055 | -1.199 |
| Notice delay | 3.57 | 97.32 | 36.8012 | 5.70299 | 27.35059 | .919 | -.232 |
| Disruptive delay | 2.68 | 99.11 | 31.7336 | 4.75623 | 23.30069 | 1.387 | 1.964 |
| Eye gaze | .00 | 86.61 | 30.3571 | 4.43247 | 21.71458 | .771 | .287 |
| Awkward silences | .00 | 67.86 | 26.7113 | 4.19508 | 20.55160 | .427 | -.884 |

Below we describe how we derived four clusters of scores (from high to low mean ratings), which are colour coded in the table above: high = red, high-mid = yellow, low-mid = green and low = blue. For a scale of 0 – 100, we can expect standard deviations between 20 – 30 and most of them are within that range with the exception of three variables, liveliness, being able to see the participants in the other rooms and ease of keeping track of the conversation; all three show a relative narrow spread, indicating a relatively high group concordance.

The distributions for all the variables were not particularly skewed to the high or low end, i.e. “skewness” for all variables fell within the recommended range of between -2 and +2.



Two variables, how lifelike participants in the other rooms looked and ease of keeping track of the conversation, showed a pronounced steepness of distribution, i.e. kurtosis was greater than +2, see figures 2.3.1.1.-1 and 2.3.1.1.-2. As can be expected from its high mean the peak for “lifelike” was towards the high end and, again reflecting its mean, for “keeping track” this was between the 50 – 60 mark.

Table 2.3.1.1-2: Paired Comparisons items in Static Condition

| | life like | sim starts | lively | see | face | track | across | close | tv panel | side | face-to-face | window | delay | disrupted delay | eye contact | silence |
|-----------------|-----------|------------|--------|------|------|-------|--------|-------|----------|------|--------------|--------|-------|-----------------|-------------|---------|
| sim starts | ns | | | | | | | | | | | | | | | |
| lively | ns | ns | | | | | | | | | | | | | | |
| see | ns | ns | ns | | | | | | | | | | | | | |
| face | .04 | .054 | .055 | ns | | | | | | | | | | | | |
| track | .003 | .026 | .029 | .072 | ns | | | | | | | | | | | |
| across | .003 | .016 | .005 | .047 | ns | ns | | | | | | | | | | |
| close | .003 | .012 | .019 | .082 | ns | ns | ns | | | | | | | | | |
| tv panel | .012 | .021 | .024 | .038 | ns | ns | ns | ns | | | | | | | | |
| side | .001 | 0 | 0 | .009 | .043 | .049 | ns | ns | ns | | | | | | | |
| face-to-face | 0 | .001 | 0 | .002 | .048 | .037 | ns | .056 | ns | ns | | | | | | |
| window | 0 | 0 | .001 | .002 | .022 | .013 | .098 | .045 | ns | ns | ns | | | | | |
| delay | 0 | 0 | 0 | .001 | .003 | .005 | .048 | ns | .076 | .099 | ns | ns | | | | |
| disrupted delay | 0 | 0 | 0 | 0 | 0 | 0 | .004 | .023 | .018 | .018 | ns | ns | ns | | | |
| eye contact | 0 | 0 | 0 | 0 | 0 | 0 | .001 | 0 | .001 | .019 | .008 | .007 | ns | ns | | |
| silence | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | .001 | .007 | .007 | ns | ns | | ns |

We performed pairwise comparisons (within subject ANOVAs) between the items as a form of discriminant analysis, which allowed us to delineate boundaries between the items in table 2.3.1.1-1. The p-values resulting from the paired comparison exercise are given in table 2.3.1.1-2. Note that non-significant p-values are marked as “ns” and p-values smaller than .10 and greater than .5 are printed in grey.

There are no significant differences between the four items “Lifelike”, “Simultaneous Starts”, “Lively” and “See remote partners”. In addition most of these items differ significantly from the items “See facial details” onwards. Thus we grouped these items together. A similar rationale was followed for the next three clusters. Although it is possible to debate the exact boundaries, based on the paired comparisons we propose to define four clusters of ratings:

1. High ratings (red)
2. High-Mid ratings (yellow)
3. Low-Mid ratings (green)
4. Low ratings (blue)

Note that it is not always possible to phrase questions in such a way that a high rating indicates a positive experience and a low rating a negative one.

1. High Ratings Static Condition

On average, participants gave the highest ratings in the static condition for their remote conversational partners being lifelike, they experienced a high number of simultaneous starts, the sessions were lively (in spite of simultaneous starts) and they could see their remote conversational partners well.

2. High-Mid Ratings Static Condition

Participants were able to see the facial expressions relatively well, could keep track of the conversation, they thought they came across well, felt close to remote partners, thought the visual representation resembled a TV-panel but were also distracted from the conversational flow by side conversations with the co-present participant, i.e. the person in the same room.

3. Low-Mid Ratings Static Condition

Poorer ratings were given to how like a face-to-face meeting the mediated session was, the configuration was not like looking through a window and seeing the remote partner. On the other hand delay was not noticed particularly and was not found to be overly disruptive.

4. Low Ratings Static Condition

The static configuration did not enable (or simulate) eye contact but in spite of simultaneous starts being rated as occurring frequently participants rated the occurrence of awkward silences low.

Figure 2.3.1.1-3 shows the means for the Static condition, as it also takes into account the spread around the means (the 95% confidence intervals) the first two variables are “swapped around”. Consistent with the tables above, the high, high-mid, low-mid and low ratings are colour coded in red, yellow, green and blue respectively. We also show which variables revealed some differences with the Orchestrated condition. This will be detailed in section 2.3.2. “Like for Like comparisons”.

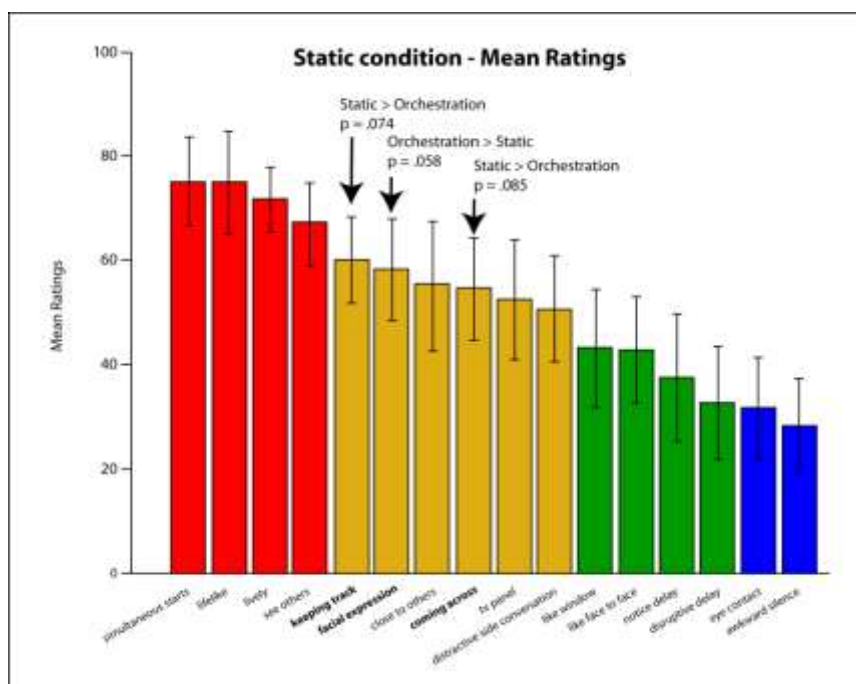


Figure 2.3.1.1-3: Means for Static Condition - 95% Confidence Intervals

2.3.3.1.2 Descriptive Statistics of the Orchestration Condition

Table 2.3.1.2.-1 shows descriptive statistics for the data stemming from the Orchestrated condition also ordered from high to low mean values.

There were three variables, simultaneous starts, how lively sessions were and the occurrence of awkward silences, that showed relative narrow standard deviations (an indication of concordance amongst the participants). Skewness and kurtosis were all within the -2 to + 2 range.

Table 2.3.1.2.-1: Orchestration Condition: Descriptive Statistics

| | Minimum | Maximum | Mean | Std. Error | Std. Deviation | Skewness | Kurtosis |
|---------------------|---------|---------|----------------|------------|----------------|----------|----------|
| See others | 21.43 | 96.43 | 73.9211 | 4.09687 | 20.07048 | -1.195 | .774 |
| Facial expressions | 4.46 | 97.32 | 72.9911 | 5.25377 | 25.73812 | -1.545 | 1.900 |
| Simultaneous starts | 28.57 | 98.21 | 72.2098 | 3.83706 | 18.79766 | -.463 | -.427 |
| Lively | 35.71 | 100.00 | 71.9494 | 3.45966 | 16.94878 | -.318 | -.501 |
| Lifelike | 16.07 | 100.00 | 70.9449 | 4.23723 | 20.75811 | -.951 | 1.066 |
| Like TV-panel | 2.68 | 100.00 | 58.2217 | 5.76762 | 28.25546 | -.543 | -.392 |
| Side conversation | 1.79 | 92.86 | 56.2128 | 4.54650 | 22.27321 | -.379 | .092 |
| Keeping track | 2.68 | 83.04 | 50.8929 | 4.84081 | 23.71504 | -.254 | -1.129 |
| Close to others | .00 | 92.86 | 47.9167 | 5.88381 | 28.82466 | -.165 | -1.087 |
| Come across | 2.68 | 83.04 | 45.4613 | 5.07328 | 24.85388 | -.363 | -1.262 |
| Disruptive delay | 3.57 | 96.43 | 42.0387 | 5.98535 | 29.32209 | .383 | -.772 |
| Notice delay | 2.68 | 91.96 | 41.2202 | 5.50438 | 26.96584 | .284 | -.892 |
| Like face to face | 2.68 | 83.04 | 36.6443 | 5.08396 | 24.90621 | .291 | -1.147 |
| Like a window | .89 | 85.71 | 34.3750 | 5.73649 | 28.10296 | .571 | -1.123 |
| Eye gaze | .00 | 81.25 | 29.9851 | 5.63718 | 27.61641 | .443 | -1.334 |
| Awkward silences | .00 | 64.29 | 27.0089 | 3.43760 | 16.84074 | .192 | -.365 |

We performed a paired comparison analysis (within subjects ANOVA's) between the items as a form of discriminant analysis, which allowed us to delineate boundaries between the items in table 2.3.1.2.-1. The p-values resulting from the paired comparison exercise are given in table 2.3.1.2.-2. Note that non-significant p-values are marked as "ns" and p-values smaller than .10 and greater than .5 are printed in grey.



Table 2.3.1.2.-1: Paired Comparisons items in the Orchestrated Condition

| | see | face | sim start | lively | lifelike | tv panel | side | track | close | across | disruptive delay | delay | f2f | window | eye |
|------------------|-------|-------|-----------|--------|----------|----------|------|-------|-------|--------|------------------|-------|-----|--------|-----|
| face | ns | | | | | | | | | | | | | | |
| sim start | ns | ns | | | | | | | | | | | | | |
| lively | ns | ns | ns | | | | | | | | | | | | |
| lifelike | ns | ns | ns | ns | | | | | | | | | | | |
| tv panel | .076 | .019 | .045 | .015 | .092 | | | | | | | | | | |
| side | .005 | 0.026 | .005 | .007 | .007 | ns | | | | | | | | | |
| track | 0.004 | 0.004 | .008 | .002 | .001 | ns | ns | | | | | | | | |
| close | 0.002 | 0.002 | .004 | .005 | 0 | ns | ns | ns | | | | | | | |
| across | 0.001 | 0.001 | .002 | .001 | .001 | ns | ns | ns | ns | | | | | | |
| disruptive delay | 0.001 | 0.001 | 0 | 0 | 0 | ns | .009 | ns | ns | ns | | | | | |
| delay | 0.004 | 0.004 | 0 | 0 | 0 | .051 | .004 | ns | ns | ns | ns | | | | |
| face to face | 0 | 0 | 0 | 0 | 0 | .015 | .009 | .022 | .039 | ns | ns | ns | | | |
| window | 0 | 0 | 0 | 0 | 0 | .005 | .006 | .027 | .025 | .097 | ns | ns | ns | | |
| eye contact | 0 | 0 | 0 | 0 | 0 | .004 | .002 | .005 | .002 | .005 | ns | ns | ns | ns | |
| silence | 0 | 0 | 0 | 0 | 0 | 0 | 0 | .002 | .002 | .006 | .049 | .055 | ns | ns | ns |

Following the same rationale as we did for grouping the questionnaire items in the Static condition, we identified three bands of ratings (High, Mid and Low) for the Orchestration condition.

1. High Ratings (red) Orchestrated Condition

Participants reported that they could see their conversational partners well, could see facial expressions very well, the conversation still suffered from simultaneous starts, but all the same the conversations were lively and remote partners seemed lifelike.

2. Mid Ratings (yellow) Orchestrated Condition

The display configuration reminded participants (somewhat) of a TV panel show, they were distracted by side conversations with the person who was co-present (in the same room), they could keep track of the

conversation, felt relatively close to remote partners, felt they came across reasonably well, found the display disruptive and they noticed the delay.

3. Low Ratings (blue) Orchestrated Condition

They did not judge the communication to be like a face to face meeting, it was not like looking through a window, eye contact was poor but there were few awkward silences.

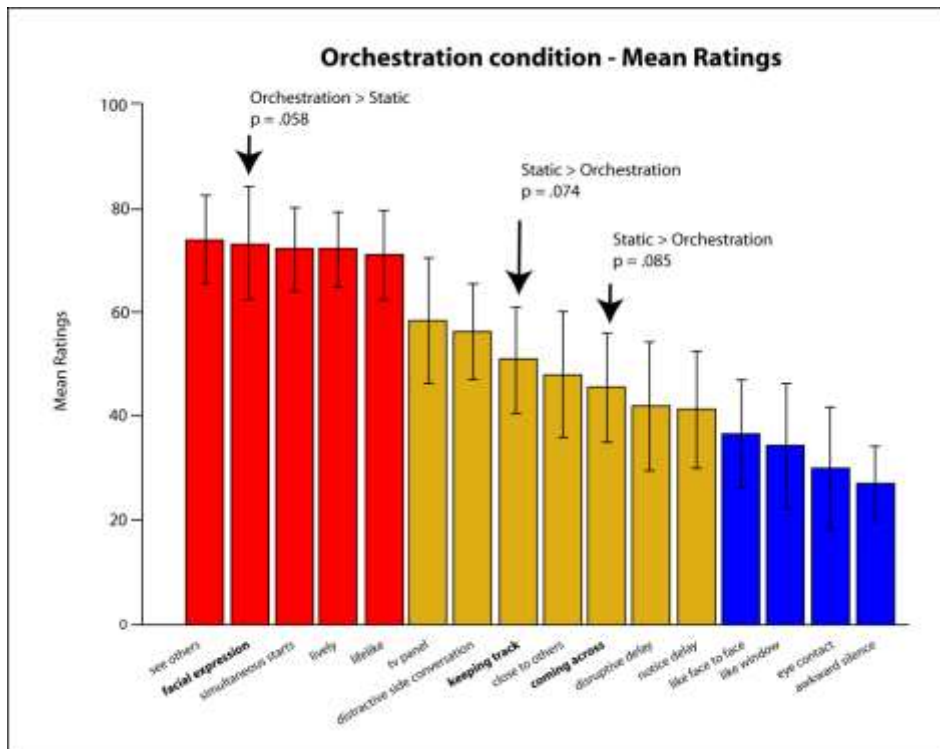


Figure 2.3.1.2.-1: Means for Orchestration Condition - 95% Confidence Intervals

Figure 2.3.1.2.-1 shows the means for the Static condition. The high, mid and low ratings are colour coded in red, yellow and blue respectively. We also show which variables revealed some differences with the Static condition. This will be detailed next in section 2.3.2. Like for Like comparisons.

2.3.3.2 Two Way ANOVAS, Like for Like comparisons

A series of two way ANOVAs was carried out comparing the survey questions (repeated measures, main effect for the two experimental conditions) in a “like for like” manner with order of presentation as a between subjects factor. As there were equal numbers of subjects per condition / cell, we used a type III error. In addition to the F-ratio and p-value we also provide the effect size: Cohen’s [1973] partial eta-squared (η_p^2).

Table 2.3.2-1: Means, Standard Deviations and p-values for significant results

| | Static Condition | | | | Orchestrated Condition | | | | p-values | | |
|--------------------|------------------|-------|--------------|-------|------------------------|-------|--------------|-------|-------------|-------------|-------------|
| | Order 1 | | Order 2 | | Order 1 | | Order 2 | | Main | Interaction | Order |
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD | | | |
| Facial Expressions | 64.14 | 21.84 | 56.25 | 21.96 | 85.94 | 10.43 | 60.04 | 30.18 | .058 | | .016 |
| Keep track | 66.15 | 14.33 | 52.53 | 19.30 | 45.09 | 25.36 | 56.70 | 21.43 | .074 | .010 | |
| Come across | 62.57 | 23.75 | 48.14 | 19.92 | 49.33 | 28.27 | 41.59 | 21.44 | .085 | | |
| Notice delay | 25.41 | 23.38 | 47.25 | 27.39 | 46.27 | 31.85 | 39.58 | 21.29 | | .035 | |
| Disruptive delay | 19.57 | 11.81 | 43.90 | 25.94 | 43.08 | 34.17 | 41 | 25.06 | | .038 | |
| Eye gaze | 38.02 | 19.91 | 22.69 | 21.48 | 45.09 | 27.71 | 14.88 | 18.14 | | | .005 |
| Like face to face | 58.33 | 19.70 | 32.89 | 21.99 | 43.82 | 27.91 | 29.46 | 20.14 | | | .006 |
| Closeness | 61.09 | 23.28 | 47.81 | 30.84 | 56.62 | 27.91 | 35.63 | 26.51 | | | .057 |
| Window | 47.84 | 23.44 | 36.44 | 27.02 | 43.08 | 34.15 | 23.30 | 16.65 | | | .092 |

Strictly speaking there were no significant main effects. Adopting a 10% significance level however there were interesting findings for three main effects, i.e. differences between the Static and Orchestrated conditions.

There were three significant Order effects ($p < .05$) and a further two where the p-value was below .10. There were three significant ($p < .05$) interactions between main effect and order.

Table 2.3.2-1 lists the means and standard deviations for nine questions for the two conditions and the two orders as well as the p-values for Main Effect (difference between conditions), Order effects and the Interactions. The results are presented in more detail below.

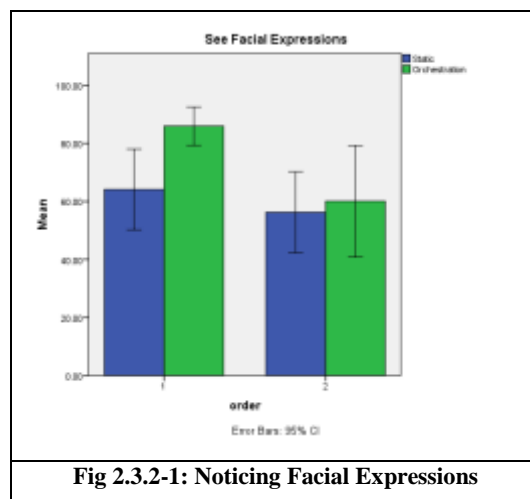
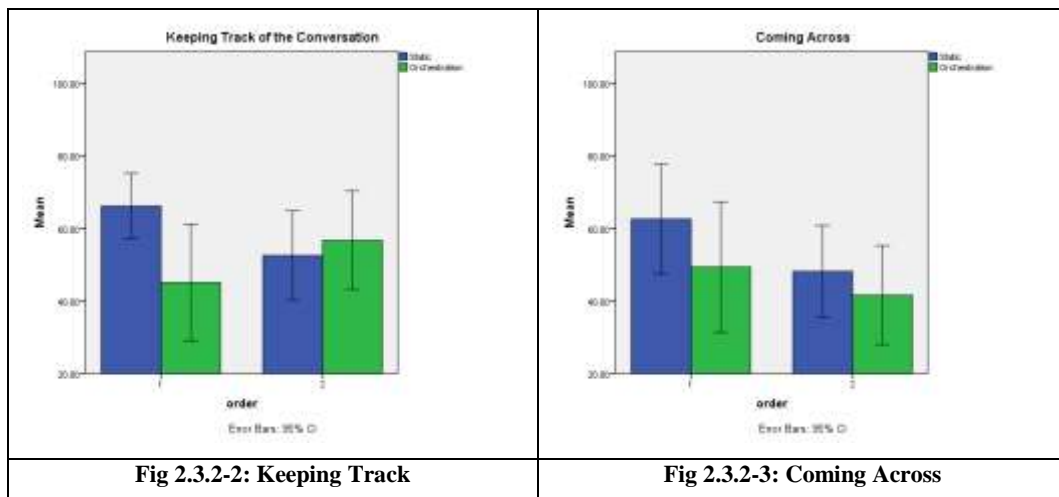


Fig 2.3.2-1: Noticing Facial Expressions

Figure 2.3.2-1 shows the means for the Static and Orchestration conditions for Order 1 (Static – Orchestration) and Order 2 (Orchestration – Static)

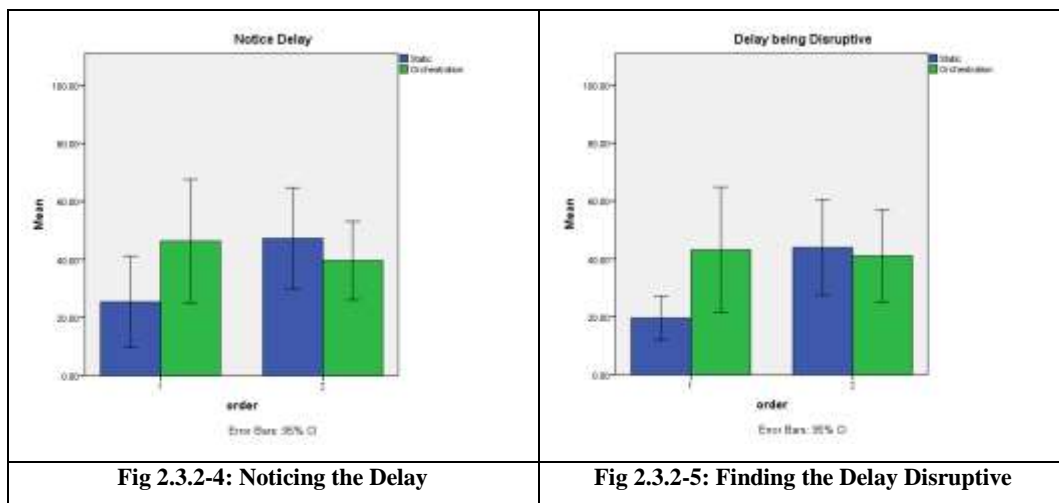
Main Effect: Static Vs. Orchestration: For being able to see “Facial Expressions” (figure @) we found a near significant Main effect, $F_{(1,22)} = 4.006$, $p = .058$, $\eta_p^2 = .154$, seeing facial expressions was rated higher in the Orchestrated condition than in the Static condition

Order Effect: There was a significant Order effect, $F_{(1,22)} = 6.864$, $p = .016$, $\eta_p^2 = .238$. The mean ratings for seeing facial expressions were highest (irrespective of condition) when the Orchestrated condition was second. Thus after participants had experienced the Static condition first they noticed that the Orchestrated condition allowed for seeing facial expressions better. When the Orchestrated condition was first (in Order 2) there was no opportunity for comparison and participants rated seeing facial expressions relatively lower (than in Order 1).

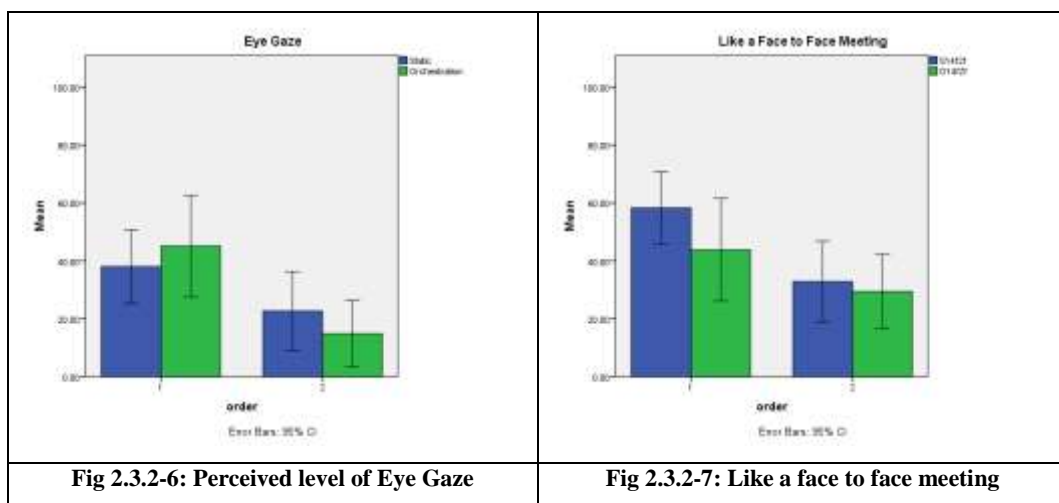


Being able to keep track of the conversation (fig. 2.3.2-2) was on average rated higher in the Static condition, although not significantly, $F_{(1,22)} = 3.522$, $p = .074$, $\eta_p^2 = .138$. There was a significant interaction, where in the second session of Order 1 (the Orchestration condition) and in the second session of Order 2 (the Static condition) lower ratings were given. Given the short duration of the task, 10 minutes, and the general liveliness of both sessions, it seems unlikely that experimental fatigue was the source of the interaction.

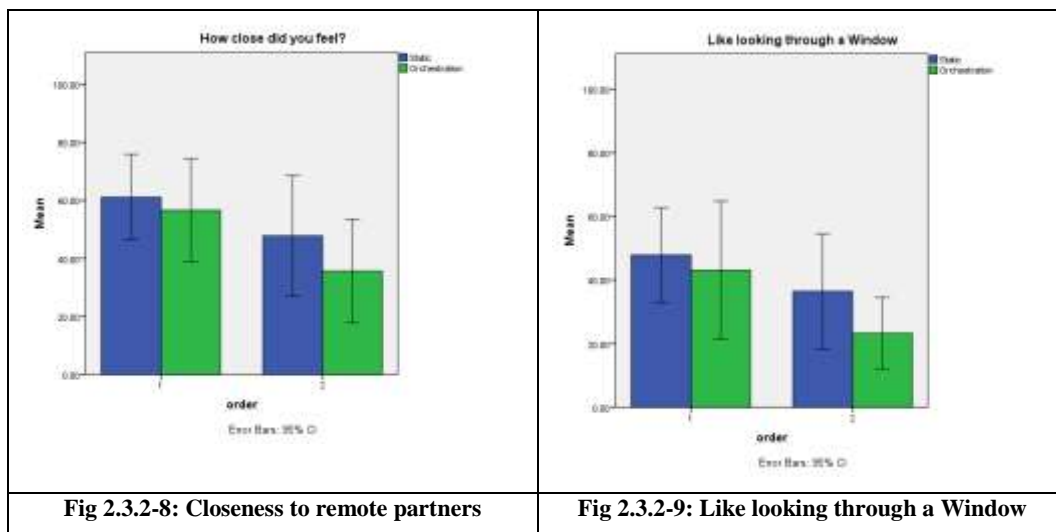
On average, participants felt they came across better in the Static condition (figure 2.3.2-3), although this did not reach significance, $F_{(1,22)} = 3.263$, $p = .085$, $\eta_p^2 = .129$.



In the second session, which in Order 1 was the Orchestrated condition and in Order 2 the Static condition, participants noticed the delay (fig 2.3.2-4) significantly more, F interaction $(_{1,21}) = 5.075$, $p = .035$, $\eta_p^2 = .195$. Similarly there was a significant interaction for finding the delay disruptive, F interaction $(_{1,22}) = 4.896$, $p = .038$, $\eta_p^2 = .182$ (fig 2.3.2-5).



There were significant Order effects for (the extent participants experienced) eye gaze (fig 2.3.2-6, $F_{(1,22)} = 9.930$, $p = .005$, $\eta_p^2 = .311$) and how like face to face meetings the mediated communication was, (fig 2.3.2-7, $F_{(1,22)} = 9.282$, $p = .006$, $\eta_p^2 = .297$). The participants in Order 1 gave higher ratings for both these variables, irrespective of condition.



There was a similar picture for Order effects for how close they felt to remote partners (fig 2.3.2-8) and how much the session was like looking through a window (fig 2.3.2-9), with participants in Order 1 (again) giving higher ratings. However the results did not quite reach significance. For Closeness the result was $F_{(1,21)} = 4.060$, $p = .057$, $\eta_p^2 = .162$ and for “Like looking through a window” this was $F_{(1,21)} = 3.125$, $p = .092$, $\eta_p^2 = .13$.

2.3.3.3 Correlations and clustering

2.3.3.3.1 Like for Like correlations

In this section we analyse where there were significant correlations between the responses to the same questions in the two experimental conditions, static and orchestration (table: 2.3.3.1.-1).

There was a highly significant correlation between the two conditions for simultaneous starts. This was also one of the questions where in both conditions the ratings were high. What is apparent from the scatter plot (figure 2.3.3.1.-1)) that indeed most participants rated the questions (almost) equally high in both conditions. However the correlation seems to derive its significance (i.e. the steepness of the regression line) from the few individuals who gave low ratings for simultaneous starts (the dots bottom left).

From the observations and the ratings it is clear that in both conditions the conversation was very lively and here the scatter plot reveals a more convincing significant correlation, a cleaner regression line than simultaneous starts (figure 2.3.3.1.-2).

The occurrence of awkward silences and having eye contact received the lowest ratings in both conditions. From the significant correlation and the scatter plots it is clear that the same participants rated this in equal measure in both conditions (figures 2.3.3.1.-3 and 2.3.3.1.-4).

Table 2.3.3.1.-1: Significant Like for Like correlations

| | r (df = 23) | p |
|---------------------|-------------|------|
| Simultaneous starts | .737 | .000 |
| Liveliness | .675 | .000 |
| Awkward silences | .503 | .012 |
| Eye gaze | .421 | .040 |
| Coming across | .382 | .065 |

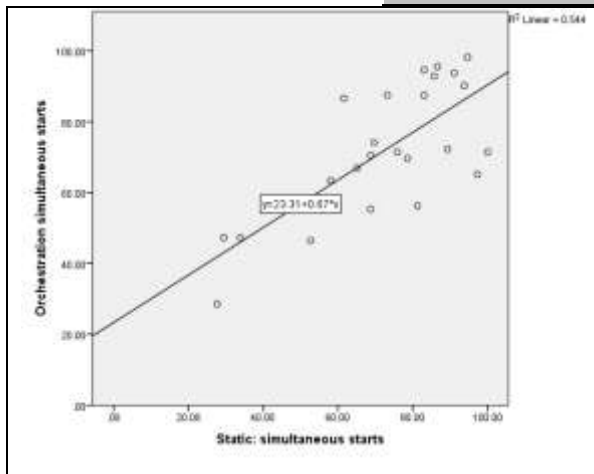


Fig 2.3.3.1.-1: Scatterplot Simultaneous Starts

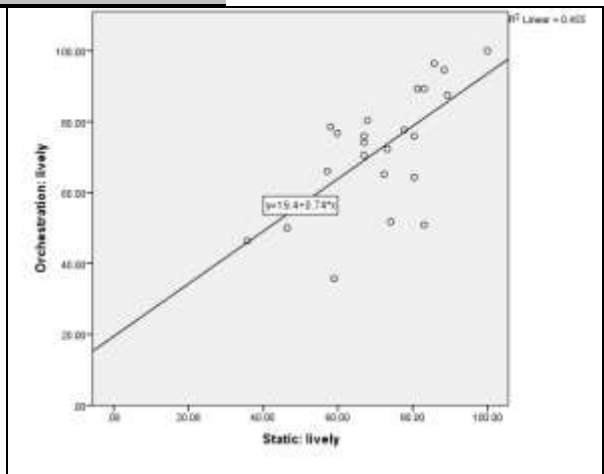


Fig 2.3.3.1.-2: Scatterplot Liveliness

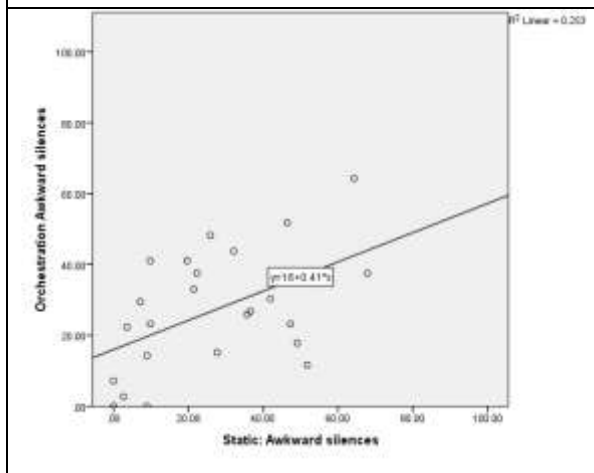


Fig 2.3.3.1.-3: Scatterplot Awkward Silences

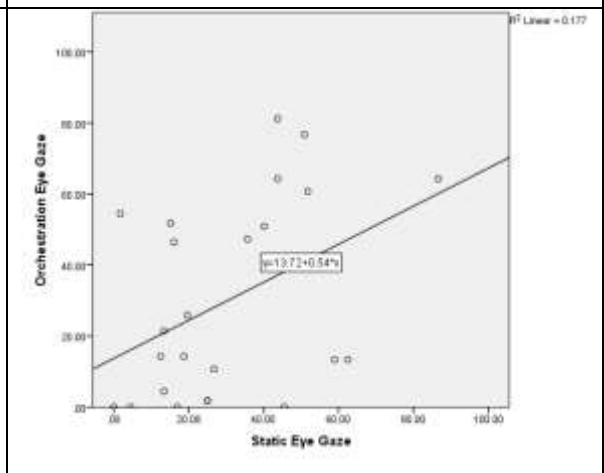


Fig 2.3.3.1.-4: Scatterplot Eye Gaze

2.3.3.3.2 Correlations Static Condition

Table 2.3.3.2.-1: Significant correlations – Static condition

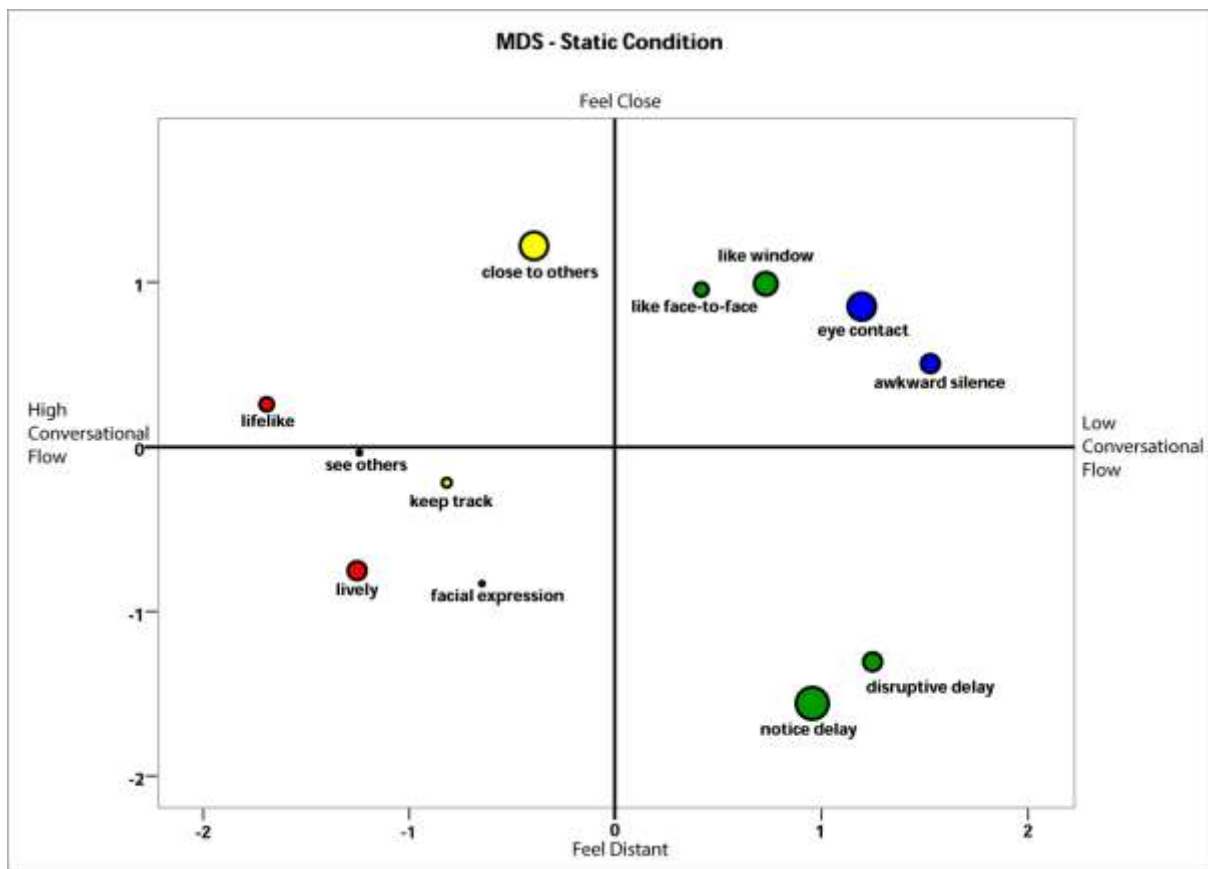
| | Window | Lifelike | Close | Face | Eye | See | Delay | Disruptive delay | Silence | Lively | Like f2f | Track |
|-------------------|----------|----------|----------|----------|----------|----------|----------|------------------|----------|----------|----------|----------|
| Lifelike | | | | | | | | | | | | |
| Close | .034 | .039 | | | | | | | | | | |
| Face | | | | | | | | | | | | |
| Eye | .000 | | .011 | | | | | | | | | |
| See | | .039 | | | | | | | | | | |
| Delay | .033 | | .007 | | .006 | | | | | | | |
| Disruptive delay | | | .008 | | .009 | | .000 | | | | | |
| Silence | .040 | | | | .003 | | .030 | | | | | |
| Lively | .003 | | | | .007 | | .037 | | .049 | | | |
| Like face to face | | | .004 | | | | .021 | .005 | | | | |
| Track | | .046 | | .022 | | | | | | | | |
| Total | 5 | 3 | 6 | 1 | 6 | 1 | 7 | 4 | 4 | 4 | 3 | 2 |

In order to explore some sort of a psychological model that would capture how participants felt and behaved in the Static Condition we carried out correlations between the responses to all 16 questions in the Static Condition.

Using this correlation matrix we derived Table 2.3.3.2.-1 which shows the p-values for the significant correlations between the questions (variables) in the static condition. Four variables did not reveal significant inter-correlations and these are omitted from the Multi Dimensional Scaling exercise described below.

The positive significant correlations are displayed in black and the significant negative ones in red. The bottom row shows for each variable how many significant correlations there were, e.g. noticing delay correlated significantly positive with two other variables and negatively with five other variables, a total of seven significant correlations.

Concentrating on the 12 variables that produced significant correlations, we applied a Multi Dimensional Scaling (MDS) analysis to visualise the inter-correlations; the size of the circles (in figure 2.3.3.2.-1) indicates how many significant correlations a variable revealed (bottom row in table 2.3.3.2.-1).



2.3.3.2.-1: MDS solution static condition (size of circles reflects the number of significant correlations)

Fig

Visualising and interpreting the relationship between 12 variables into a two-dimensional space, as is suitable for a written report, is a very useful but also somewhat artificial exercise. Here we depart from the conventional style of a results section in which (in this case statistical) findings are (drily) listed and interpretation is reserved for the Discussion.

The size of the circles in figure 2.3.3.2.-1 reflects the number of correlations with other variables and variables such as “noticing the delay”, “feeling close to others” and “the subjective experience of having eye contact” may have more predictive power.

In this plot the Y-axis is interpreted to reflect levels of closeness. Feeling close (Telepresence) is of course an important one, but when video conferencing works well, i.e. people feel close, the experience has been described as looking through a window (as during a prison visit) and there is good level of experience of eye contact (e.g. see positive correlation between eye contact and closeness) . Delay on the other end of the continuum has been seen traditionally as an obstacle to feeling close (Telepresence).

The X-axis is interpreted as reflecting the levels of conversational flow, e.g. finding the discussion lively, being aware of others, keeping track of the conversation, using facial expressions as conversational feedback/backchannel. Awkward silences (an interesting positive correlation with eye contact) and finding delay disruptive (negative correlation with eye contact) signify a low conversational flow.

Incidentally the colour coding (the order of high to low means of the variables) has been maintained. It is not unusual for a one dimensional (MDS) solution to reflect this order; in this case a perpendicular projection onto the X-axis approximates such a one dimensional solution.

2.3.3.3.3 Correlations Orchestrated Condition

Repeating this exercise for the Orchestrated condition, we show the significant p-values for the correlation matrix for the questionnaire in the Orchestrated condition in table 2.3.3.3-1. The p-values of the significant positive correlations are in black and the negative ones in red.

In the Orchestrated condition all 16 variables produced at least one significant correlation although this did not exceed four significant correlations.

The MDS plot (fig. 2.3.3.3-1) can be interpreted along the (same naming of the) axes. Delay takes up the same low position in the quadrant bottom right. Eye contact and awkward silence are in a similar position. At the high end of conversational flow there are “lively” and “seeing others”. However there are some shifts. Most notably seeing facial expressions seems to contribute to the conversational flow.

Table @: 2.3.3.3-1: Significant correlations – Orchestrated condition

| | Window | Like TV | Lifelike | Close | Face | Eye | Side | See | Delay | Disruptive | Simstart | Silence | Lively | Face to face | Track | Across | Total |
|------------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|------------|----------|----------|----------|--------------|----------|--------|-------|
| Like tv | | | | | | | | | | | | | | | | | |
| Lifelike | | | | | | | | | | | | | | | | | |
| Close | .008 | | .036 | | | | | | | | | | | | | | |
| Face | .032 | .037 | | | | | | | | | | | | | | | |
| Eye | | | .003 | | | | | | | | | | | | | | |
| Side conv. | | | | | | | | | | | | | | | | | |
| See | | .025 | | | | | | | | | | | | | | | |
| Delay | | | .032 | | | | .004 | | | | | | | | | | |
| Disruptive delay | | | .022 | | | | .003 | | .000 | | | | | | | | |
| Sim. start | | | | | .037 | | | | | | | | | | | | |
| Silence | .024 | | | | | | | | | | | | | | | | |
| Lively | | .030 | | | .003 | | | | | | .000 | | | | | | |
| Face to face | .008 | | | .004 | | .000 | | | | | | | | | | | |
| Track | | | | | | | | .049 | | | | .031 | | | | | |
| Across | | | | | | .003 | | | | | | .015 | | | .018 | .043 | |
| Total | 3 | 3 | 4 | 4 | 4 | 3 | 2 | 2 | 3 | 3 | 4 | 1 | 3 | 4 | 4 | | |

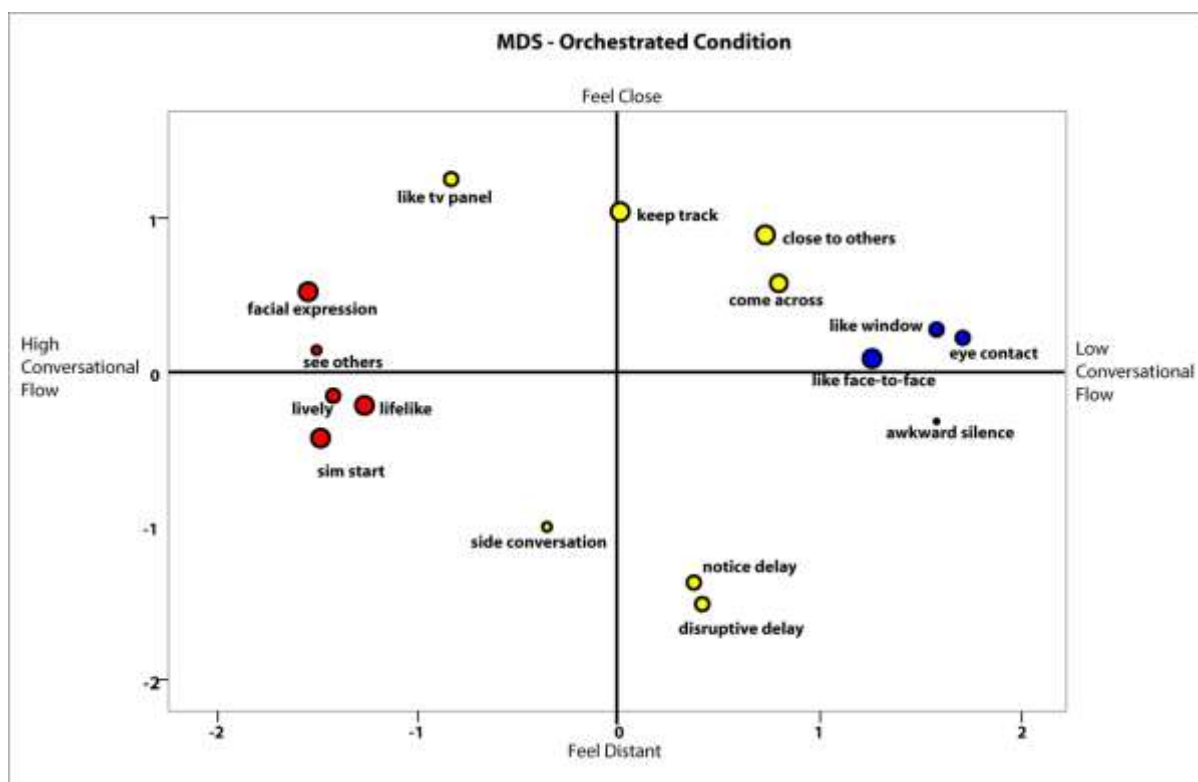


Fig 2.3.3.3.-1: MDS solution Orchestrated condition (size of circles reflects the number of significant correlations)

2.4 Discussion

Using a laboratory environment, we compared the social communication between three “living rooms” either using one static camera (Static Condition) or using three cameras in an Orchestrated fashion (Orchestrated Condition). In each room there were two participants.

We took a number of complementary measures to evaluate the differential effects of the two experimental conditions: group interviews after an experimental session, questionnaires and automated logging of vocal activity in the three experimental rooms.

Interviews

The interviews with the groups of six participants after the experiments identified a number of themes with potential feedback into future iterations of an orchestration logic.

Whilst close up shots enhanced Telepresence (compared to wide shots), it also interrupted a consistent visual as well as conversational flow of the communication, i.e. it gave participants the impression that they were in separate spaces. Some found discontinuity in display, the cutting of shots, itself quite distracting whilst others even found the errors in orchestration (when often it wasn't showing who was talking) helped in creating a sense of equality by showing more of the non-speakers. As such orchestration needs to take into account how the screen layout can affect the segmentation of the personal space. For example these comments would suggest using a split screen of multiple shots might have a more unifying effect on a group rather space than full screen cutting of close up shots which creates a more private one-one-effect. However, the split

screen layout resulted in much smaller images and on occasion made it difficult to identify a speaker, in particular when more than one person was talking simultaneously.

Orchestration aims to create an immersive experience but communicating across TV screens and edited vision mixes, inevitably introduces a different dynamic to interactions when compared to talking together face to face in a group. In particular there was no feedback about whether you were seen or heard. Some (more robust) participants were able to adapt to lacking the feedback whilst others felt excluded.

Seeing facial expressions better, full screen orchestration seemed to provide a more intimate experience compared to the more formal one offered by a static split screen composition. The static screens were better for getting things done it was felt, but the orchestrated close ups made it more fun and intimate. This builds on the idea of the use of close up shots creating a sense of a series of one on one type conversation.

This would imply orchestration, in conjunction with control of the screen layout, could “choose” different screen compositions and different orchestrating styles according to the context of the interaction e.g. formal class room or simple supervision, or formal meeting versus friends getting together.

The current implementation of the orchestrator relies upon voice activity primarily as a cue, with these (directional) audio cues being further interpreted into conversational turns. However, rather than merely using the audio channel as a source to extract cues from, orchestration maybe needs to include the audio as an integral part of the experience and of the systems understanding of the context of the conversation, e.g. whether people know each other (and their conversational partners’ voices) or not.

For example, if people already know each other then it is easier for them to notice recognize each other from the audio. This would mean that it is less crucial for the correct speaker to be shown on screen at all times when participants in a video conference know each other compared to when they do not. This additional information contained in the audio channel may be able to mitigate the segmentation of the space resulting from the use of close ups described earlier if the people know each other and allow for the benefit of the intimacy which come with close up shots without compromising too much on the group experience. How well participants know each other then needs to become an input into the orchestrator knowledge. This input, i.e. knowledge of how well people know each other, would be an input that would need to come from a source outside of the immediate space of the interaction and using this input would mean orchestration would be varying the experience according to the communication context.

This opens the door to varying orchestration, shot choices and screen layout, according to how well the people involved know each other. For example, if 1 person out of 6 happens to be on one Vclient on his own then he might receive a split screen showing 5 small close ups of the other participants to allow him/her to orientate themselves better in the communication whilst the other would receive full screen orchestrated feeds. Also, when a person joining a new conference might be unknown to the other participants then a period of close ups of the other participants might help the process of introduction before settling down on a tiled layout such as the one used in the Viewmode experiments of Chapter 1.

However, this might be a very difficult problem to resolve accurately. At all times we must be aware of the cost of a system getting it wrong Vs. the value of a system getting it right. In addition display modes benefit from consistency and continuity. Too many changes in display during a conversation might be disconcerting for participants.

The difficulty of using audio as a main driver for orchestration becomes compounded when people interrupt each other, in cases of fast (overlapping) turn taking (cross talk). In particular, the existing orchestration logic had built into it a limiter which prevented an edit being made before an existing shot has been on screen for eg. 2 secs. This breaks down when conversation turns are shorter than 2 secs.

However, this in itself could possible provide an additional cue into screen layout choices. For example, if there were a continued period of cross talk then the orchestrator could maybe choose a layout that best represented the two relevant people to the other.



Questionnaires

For both conditions there was an identical set of 16 questions, asking about how close participants felt to participants in the other rooms, how they felt they came across to the participants in the other rooms, conversely, how distracting side conversations were with the co-present participant; to what extent they felt there was eye contact; how much the communication was like looking through a window, like a TV-panel (game-show), like a face to face meeting; conversational parameters such as how lively the conversation was, how often there were awkward silences, how often there were simultaneous starts, how easy it was to keep track of the conversation. In addition the questionnaire asked about effects of latency.

In the static condition the responses could be group in four bands of scores:

High ratings were given for remote conversational partners being lifelike, they experienced a high number of simultaneous starts, the sessions were lively and they could see their remote conversational partners well. The band of High-Mid ratings indicated that participants were able to see the facial expressions relatively well, could keep track of the conversation, they thought they came across well, felt close to remote partners, thought the visual representation resembled a TV-panel but were also distracted from the conversational flow by side conversations with the co-present participant, i.e. the person in the same room. Poorer ratings were given to how like a face-to-face meeting the mediated session was; the configuration was not like looking through a window and seeing the remote partner. On the other hand delay was not noticed particularly and was not found to be overly disruptive. Low ratings were given for eye contact and participants rated the occurrence of awkward silences low.

In the Orchestrated condition participants also reported that they could see their conversational partners well, and in addition they could see facial expressions very well, the conversation still suffered from simultaneous starts, but all the same the conversations were lively and remote partners seemed lifelike. Mid ratings indicated that the display configuration reminded participants (somewhat) of a TV panel show, they were distracted by side conversations with the person who was co-present (in the same room), they could keep track of the conversation, felt relatively close to remote partners, felt they came across reasonably well, found the display disruptive and they noticed the delay. They did not judge the communication to be like a face to face meeting, it was not like looking through a window, eye contact was poor but there were few awkward silences.

The differences between the Static and Orchestrated conditions were not profound. There was a nearly significant effect for being able to see “Facial Expressions”, seeing facial expressions was rated higher in the Orchestrated condition than in the Static condition. In the static condition it was marginally easier to keep track of the conversation and there was some suggestion participants thought they came across better in the static condition.

Participants took part in the experiments in one of two orders; in order 1, they started in the static condition and then proceeded into the orchestrated condition. In order 2 this was the other way around. There were some order effects but they had no relevance to the research questions.

Cluster analysis based on correlations allowed us to describe a two dimensional space where one axis signified closeness and lack thereof, the latter mainly affected by latency. The other axis signified the level of conversational flow. In the orchestrated condition being able to see facial expressions enabled both a sense of closeness and a mechanism to function as a conversational backchannel, i.e. it helped the conversational flow.

Concluding the questionnaire section, it is fair to say that there were no dramatic differences between the two ViewModes. In the static condition there seemed to be more group awareness facilitating participants to keep track of the conversation. The most striking difference resided in the being able to see facial expressions in the Orchestrated condition, enhancing a level of Telepresence of the active speaker.

Automated logging

The system logged a number of conversational parameters as well as editing (orchestration) actions undertaken by the system (in the orchestrated condition only), based on detecting the voice activities of the participants. The logging data of the latter are not clear cut to interpret but the conversational parameters seem to point (on occasion) to a livelier conversation in the orchestrated condition.

Although there was a tendency for there to be a higher number of conversational turns in the orchestrated condition. The number of short turns and interruptions in the orchestrated condition were significantly higher. Herb Clarke argues that a high frequency of short turns and interruptions is indicative of a healthy and lively (group) conversation. Although consistent with liveliness of the conversation, these results could also signify an attempt by the group to keep track of the conversation.

On the other hand there was no significant difference in turn duration between the two conditions. Poor Telepresence has been associated with longer turn duration. Although the duration of the orchestrated shots is significantly lower than in the static condition (although in the static condition this was of course not acted upon).

2.5 Conclusion

From this and the previous experiment it seems clear that with regards to full screen display of partners in distributed video mediated group conversations, there is no “one size fits all”. Whilst at the small computer screen, the full screen view mode in a lively group conversation detracted from group Telepresence. In a living room scenario the Orchestrated full screen view mode on a large TV enhanced Individual Telepresence dramatically but this somehow became a feature of the group communication. In the Static view mode it was not possible to see one self which may have contributed to the lack of significant differences between the two view modes.

2.6 References for both View Mode experiments

1. Clark-Carter, D. *Doing Quantitative Psychological Research, from Design to Report*, Taylor & Francis, 1997
2. Clark, H.H. and Brennan, S *Grounding in communication*. Resnick, L.B. Levine, J. and Teasley, S.D. *Perspectives on Socially Shared Cognition*, APA Press, Washington, (1991).
3. Clark, H. H., *Using language*. New York: Cambridge University Press, (1996).
4. Cohen, J. Eta-squared and partial eta-squared in fixed factor anova designs. *Educational and Psychological Measurement*, 33, 1(1973), 107-112
5. Geelhoed, Erik, Aaron Parker, Damien J. Williams & Martin Groen. *Effects of Latency on Telepresence*. (2009) HP labs technical report: HPL-2009-120.
6. O’Conaill, B., Whittaker, S. and Wilbur, S. *Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication*. *Human-Computer Interaction*, 8, (1993), 389–428.
7. Iekin, S., K Wac, M Fiedler, L Janowski. *Factors influencing quality of experience of commonly used mobile applications*. *Communications Magazine, IEEE* (Volume:50 , Issue: 4), April 2012
8. Judge, T.K., C Neustaedter. *Sharing conversation and sharing life: video conferencing in the home*. SIGCHI Conference on Human Factors 2010 - dl.acm.org
9. Kegel, I. *Description of Proof of Concept Components*, December 2012
10. Schreiber, M. T Engelmann. *Knowledge and information awareness for initiating transactive memory system processes of computer-supported collaborating ad hoc groups*, - *Computers in Human Behavior*, 2010 – Elsevier.



-
11. Sellen, A.J. Speech patterns in video-mediated communication. In Companion Proc. of CHI'92 Human Factors in Computing Systems, ACM Press, (1992), 49-59.
 12. Sellen, A.J. Remote conversations: The effects of mediating talk with technology. *Human-Computer Interaction*, 10, (1995), 401-444.
 13. Stone, H., J.Sidel, S.Oliver, A. Woolsey & R.C. Singleton. (1974). Sensory Evaluation by Quantitative Descriptive Analysis. *Food Technology*, Nov 1974, 24-34.
 14. Whittaker, S. Rethinking video as a technology for interpersonal communication: Theory and design implications. (1995) *International Journal of Human-Computer Studies*, 42(5), 501-529.
 15. Wilson, M. and Wilson, T.P. An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review*, 12, 6 (2005), 957-968.

3 Sensing the Audience

Chen Wang¹, Erik Geelhoed², Pablo Cesar¹, Phil Stenton² & Ian Biscoe²

¹ Centrum voor Wiskunde&Informatica, Amsterdam

² Falmouth University

3.1 Abstract

Psychophysiological measurement has the potential to play an important role in audience research. Currently, such research is still in its infancy and usually involves collecting data in the *laboratory*, where during each experimental session *one* individual watches a video *recording* of a performance. We extend the experimental paradigm by *simultaneously* measuring Galvanic Skin Response (GSR) of a *group* of participants during a *live* performance. GSR data were synchronized with video footage of performers and audience. In conjunction with questionnaire data, this enabled us to identify a strongly correlated main group of participants, describe the nature of their theatre experience and map out a minute-by-minute unfolding of the performance in terms of psycho-physiological engagement. The benefits of our approach are twofold. It provides a robust and accurate mechanism for assessing a performance. Moreover, our infrastructure can enable, in the future, real-time feedback from remote audiences for online performances. We are currently scaling up the system allowing for simultaneous GSR measurement of larger audiences.

3.2 Introduction

In our work we are exploring how current video-mediated technologies can be used, and extended, for supporting novel interactive performances. In particular, our final goal is to provide the infrastructural components and support, so performing artists can reach a wider remote audience with their productions, but still maintaining the close relationship between the actor and the audience.

Live performances of big productions are already streamed to cinemas and to homes [7]. However, these present limited opportunities for audience feedback or interaction. As said by one performer “it feels more like a rehearsal than a play, as I cannot see the reactions from the audience”.

Based on these limitations, the primary motivation for the current study, and our research question, is to explore the viability of using Galvanic Skin Response (GSR) to monitor co-located and remote audience feedback during a live performance. We took GSR measurements of 15 people watching a live theatre performance simultaneously. The readings were synchronized with video recordings of the performance and the audience. The audience filled out questionnaires aimed to evaluate the emotions that the performance evoked. This resulted in a high volume of useful data of around 1680 data points for each participant.

Results indicate that our approach – gathering GSR data during the play - is valid, as such data accurately reflects the engagement of the audience members. Moreover, it proves to be a useful tool for temporally unfolding the experience of the public, as the reactions of the public can be mapped to specific events during the play. In principle, we can conclude that our solution of using GSR data for monitoring audience feedback is novel and very valuable.

This section of the report is structured as follows. First, we highlight the novelty of our approach, when compared to prior works. Then, we describe a first relevant pilot study and the methodology we use in the actual study, followed by the results that support our hypothesis. Finally, the results are analysed and discussed, focusing on the most important implications for next-generation video-mediated performances.

3.3 Related Work

Jennifer Radbourne details the importance of audience feedback through an extensive literature review and in-depth interviews [10]. The study justifies our hypothesis, as it shows that audiences are not primarily passive and that gauging the audience experience might provide an important measure of quality in the performing arts.

GSR measures excitation of the sympathetic nervous system and combined with other types of physiological and neurological, as well as self-report measures, have been applied in many areas of research, e.g. psychology, medical, gaming and education. Pejman et al. have explored how GSR data can be used for improving game design [2]; GSR sensors were also extensively used in research with hyperkinetic children [3]. For the purpose of this paper, we narrow the scope to audience feedback and how such measures open up new possibilities for interaction.

There are few studies using physiological measurements for learning about audience engagement per se (e.g. [1],[13]). There seems to be more interest in applying sensor feedback creatively, e.g. to influence the outcome of a movie [4]. The most related work is the one from Celine Latulipe [6] who provides an extensive overview of research in this area. Her work draws on the empirical and theoretical work of Peter Lang [5], who describes a two dimensional space of different emotional states where one dimension runs from low to high (GSR) arousal and the other from low to high pleasure. Latulipe and colleagues explore how bio-feedback (in particular GSR) can be used to provide real-time visual feedback to performers. In interviews with dance and theatre experts the notion of “valence” is introduced, i.e. how GSR arousal validates audience response.

Nevertheless, all these studies involve experimental sessions in which a single person watches a video recording of a performance. We instead took GSR measurement into the “field”, the natural habitat, and took collect the data of the audience simultaneously during the play. As such, we believe that our system supports ecological validity much better than the previous laboratory experiments.

3.4 Pilot Study

Before the real experiment, we conducted a pilot study with the Miracle theatre. The purpose of the pilot study was to get some insights about how to simultaneously gather GSR responses from several people during a live theatre performance. Four participants, two of them a couple, were invited to wear the GSR sensor during the performance. Three participants had a relaxed day and the other one did hard house painting work during the daytime. The pilot study results displayed a similar pattern among the three (relaxed) participants’ GSR data distribution, in particular the couple showed an unusual strong correlation ($r = 0.89$, $p < .01$). This data further encouraged us, as it looked as if reactions of people during the play could be clustered into well-differentiated patterns.

3.5 Method

Seven females (mean age 28.29) and eight males (mean age 23.13) formed the audience for a 28 minute theatre performance. Their GSR was measured every second throughout, resulting in 1680 data points for each participant. Actors devised and performed a comedy that was aimed at audience participation and produced occasional “shocks” (e.g. a popping balloon) to elicit the occurrence of GSR spikes during the performance (fig. 3.1).

The GSR measurement system consisted of 15 GSR sensors. Groups of five sensors were each connected to one of three Arduino UNO boards (sample rate 1Hz). Xbee RF modules were used to create a wireless network such that the GSR data were sent directly to a laptop. This ensured the synchronization of all

GSR readings. Cameras recorded the audience and the performance. Video streams were synchronized (post production) with GSR data.

Before the performance, participants filled out a short questionnaire asking about the type and intensity of the emotions they had experienced during the day. Afterwards participants filled out a similar questionnaire asking about emotions experienced during the play. The questionnaires were in the form of graphic rating scale [12] and measured 100mm. Participants were asked to make a mark between two extremes, i.e. between “not at all” and “very much”.

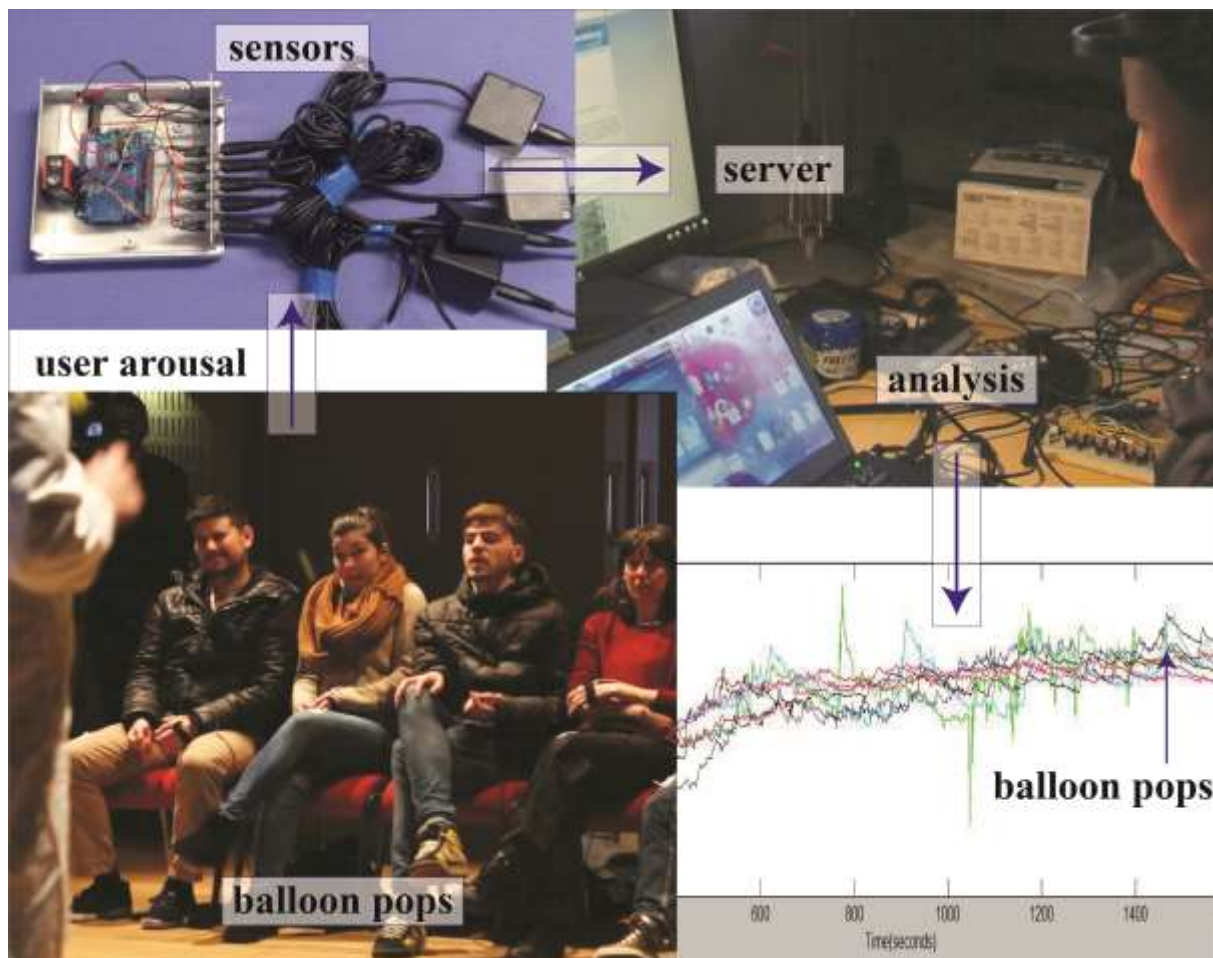


Figure 3.1: GSR system

Participants were seated in one row with three sections of five seats each, arranged in a semi-circle around the stage. GSR modules were attached to the palm of the right hand. Before the performance started, participants took part in a meditation exercise to establish a baseline GSR level.

Questionnaires were analysed using Analysis of Variance (ANOVA) and correlations. The synchronized GSR and video streams enabled us to relate events during the performance to corresponding GSR readings. GSR readings were analysed using the Multidimensional Scaling (MDS) method [9]. Correlations and ANOVA had some limitations to do a complete interpretation of the readings. They are fairly suitable if the audience is being treated as a whole, but they cannot properly explain relationships – similarity and dissimilarity - between objects in a multi-dimensional space. In our case, we were interested in understanding the relationships between 15 objects (each audience member) GSR responses’. We calculated the dissimilarities between the objects using Pearson Correlation Coefficients, and two-dimensional scaling

was chosen for scaling. After 30 iterations, the final configuration graphs were achieved and Kruskal's stress reported in the results.

3.6 Results

3.6.1 Audience clustering

A MDS solution (fig. 3.2), based on correlations of GSR readings between audience members, shows how ten participants correlated closely (on average $r = .86$), showing an initial rise in GSR followed by a flattening towards the end of the performance (inset in fig. 3.2). In this plot the Kruskal's stress value is 0.06 (less than 0.10), indicating that the configuration of the 15 participants' GSR readings can be considered as reliable. Questionnaire results and brief interviews after the performance indicated that this group had been deeply engaged in the performance, reporting high levels of enjoyment.

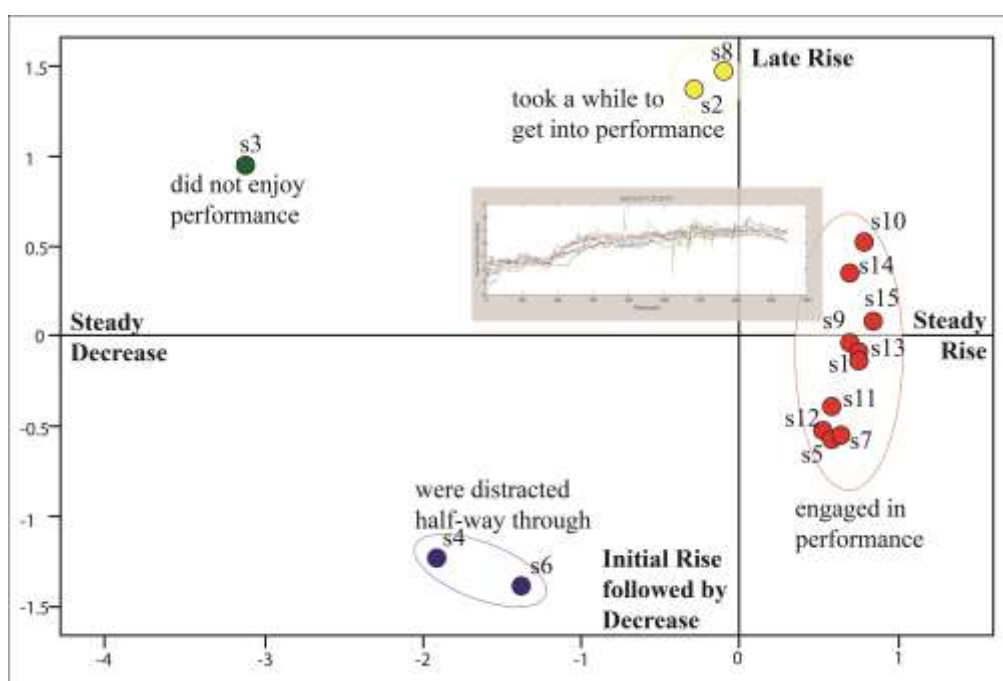


Figure 3.2: MDS Audience clustering based on GSR data

Five participants displayed different patterns. Two showed an initial rise in GSR followed by a decrease, i.e. after an initial engagement with the performance their attention waned; for one this was related to receiving sad news during the day. Two showed an initial lack of rise in GSR followed by an increase; they reported to be confused initially by the purpose of the play and as such it took them awhile to get into the performance. One participant displayed a consistent drop in GSR and reported not liking the performance. These characteristics enabled us to label the extremes of the X and Y-axes.

3.6.2 Unfolding of the performance

For each minute, the GSR readings were averaged for each participant. Here MDS (Kruskal's stress: 0.05) yielded an almost chronological minute by minute unfolding of the play (anti-clockwise in fig. 3.3) up to minute 19. Using the video footage we were able to identify the clusters based on the content of the performance. Thus, initially the GSR readings are low (minute 1) followed by a steady rise (minute 2 – 19)

after which the intensity of the GSR flattens (minute 20 – 28). The first part of the performance (minute 2 – 16, in red in fig. 3) built up to an active and physical participation during which the participants were asked to raise either their left or their right leg in response to (silly) questions by the actors. In minute 17 – 19 (in green) the results of a competition were revealed, where the relatively higher GSR readings might indicate anticipation. After that the audience was not required to interact as actively as they listened to a trumpet player (dark blue) and watched a juggling act (yellow). The Y-axis reflected levels of GSR intensity and the X-axis ran between low and high audience participation.

Spikes were identified that corresponded to the intended “shocks”, e.g. balloon popping, the sudden sound of a (badly played) trumpet.

The minute average GSR readings during this comic play correlated positively with participants being (very) cheerful (on average $r = .619$) and correlated negatively with participants being sad (on average $r = -.595$) at different stages of the performance, in particular from minute 16 onwards the average GSR readings showed strong correlations with audience’s “cheerful” ratings.

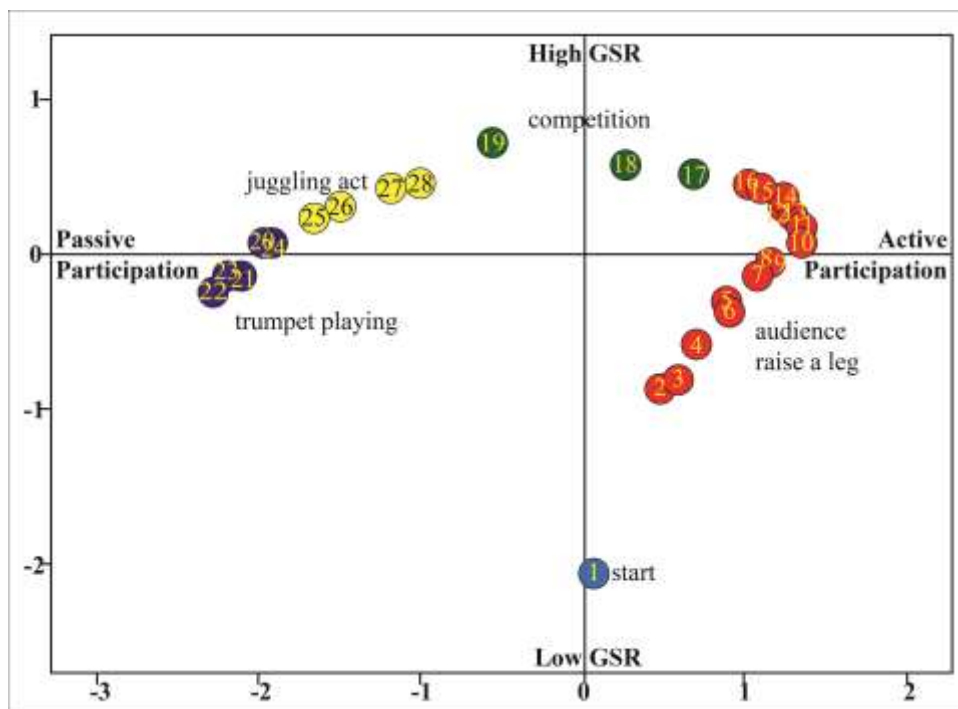


Figure 3.3: MDS minute by minute unfolding of the performance

3.6.3 Pre- and Post- Performance Questionnaires

Table 3.1 summarizes the significant differences between pre- (asking about their experiences during the day) and post-performance questionnaires. The ratings were given on a scale between “not at all” (= 0) and “very much” (=100). Thus participants rated that during the day, on occasion, they had a laugh with a mean intensity of 45 (Mean pre in Table 3.1) and they reported that the intensity of laughter generated by the performance was rated on average as 68.5 (Mean post), resulting in a significant difference, $F(1,13) = 14.68, p = .002$. Similarly, for cheerfulness, the difference between pre-and post- ratings was significant, $F(1,14) = 7.12, p = .018$. On average, participants had a reasonably cheerful day (Mean = 55) but these ratings increased to on average 74.5 after the performance. Lastly, although participants did not have a particularly sad day (with the exception of one participant) yielding a mean of 35, this was significantly reduced to a mean of 11.4 after the performance $F(1,14) = 5.82, p = .03$. There were also significant effects for gender and whether a participant

knew another participant sitting in their row or not. However, due to the low numbers in each “cell”, we refrain from reporting these in this report.

Table 3.1: Significant differences between pre-and post- questionnaires

| Item | p | Mean pre | Mean post |
|----------|------|----------|-----------|
| Laugh | .002 | 45 | 68.5 |
| Cheerful | .018 | 55 | 74.5 |
| Sad | .03 | 35 | 11.4 |

3.7 Discussion

This paper describes an audience experience during a live theatre performance using a system to measure GSR of 15 people simultaneously. Psychophysiological measurement in audience research is usually carried out in a laboratory, where during an experimental session one individual watches a recording of a performance. There are advantages to such studies as a range of physiological and neurological sensors can be used concurrently. However, being part of an audience is a group experience and it might not be straightforward to extrapolate from an individual’s experience watching a recording to a larger audience watching a live performance. As such, we believe that our system supports ecological validity much better than laboratory experiments. It is not just an innovative contribution to audience research methods that makes this study of interest. We found that for most participants there was an unexpected and unusually high level of physiological closeness (GSR), which, to our knowledge, has not been reported before.

Analysing the GSR data in conjunction with synchronized video recordings provided additional insights, e.g. we were able to link spikes in GSR to shock-effects during the performance. This validates the robustness and accuracy of our measurements. However, more interesting are the general (smoothed) shapes of audience engagement. We found that GSR readings of most of the audience followed a curve where in the initials stages readings were low and as the play progressed this increased steadily, reflecting an increase in engagement with the play across time. Returning to the notion of valence [6], low GSR does not necessarily imply a negative audience judgment, for peaks to happen troughs are essential, but it is informative to evaluate the overall shape of the response, as exemplified by the steady decrease in GSR of one participant who was not engaged by the play. In addition we could place the performance in a two dimensional space, not unlike Lang’s [5] where one dimensions runs from low to high arousal and the other from low to high audience participation. Embedding the audience experience in how their day had been also showed how a performance can lift an audience out of the ordinary. Cheerfulness and enjoyment were the main (more) emotional components linked to the GSR data.

3.8 Conclusion and Future work

We are currently in the process of scaling up the system, taking advantage of small form-factor developments in wireless technology and GSR measurement. Heart rate and blood pressure might make useful additions.

There are some near future applications of a relatively low-cost system such as the one described here. In a “next-bench” type of fashion it can be used for further audience research. It makes it feasible for theatre companies to receive detailed (and time-stamped) early feedback during try-outs, to evaluate what works well and what does not or rather identify where audience engagement wanes.

Our research focuses in providing mechanisms, so remote audiences can interact during the streaming of a live performance. Wearable physiological sensors have the potential to open a whole array of creative solutions to suite this aim, e.g. the aggregate response of those who are deeply engaged in a performance can be used to provide visual, auditory or even haptic feedback [8] [11]to performers.

References

1. S. Bardzell, J. Bardzell, and T. Pace. Understanding affective interaction: Emotion, engagement, and internet videos. Proc. of the IEEE International Conference on Affective Computing and Intelligent Interaction (2009), pp. 1-8
2. P. Mirza-Babaei, L. E. Nacke, J. Gregory, N. Collins, and G. Fitzpatrick. How does it play better? exploring user testing and biometric storyboards in games user research. Proceedings of CHI (2013), pp. 1499-1508.
3. James E. Hastings, Russell A. Barkley. A Review of Psychophysiological Research with Hyperkinetic Children. Journal of Abnormal Child Psychology (1978), 6(4): 413-447
4. C. Grant. Many Worlds: The movie that watches its audience. BBC News (2013) <http://www.bbc.co.uk/news/technology-21429437>
5. P. Lang. The emotion probe: Studies of motivation and attention. American Psychologist (1995). 50(5): 372–385.
6. C. Latulipe, E.A. Carroll, and D. Lottridge. Love, hate, arousal and engagement: exploring audience responses to performing arts. Proceedings of CHI (2011), pp. 1845-1854
7. National Theatre Live: <http://www.nationaltheatre.org.uk/>
8. S. Baurley P. Brock, E. Geelhoed, and A. Moore. Communication-Wear: User Feedback as Part of a Co-Design Process. Haptic and Audio Interaction Design (2007), pp. 56-68.
9. Gerry P. Quinn, Michael J. Keough. Experimental Design and Data Analysis for Biologists (2002).
10. J. Radbourne, K. Johanson, H. Glow, and T. White. Audience experience: measuring the quality in the performing arts. International Journal of Arts Management, (2009), 11(3): 16-29.
11. S. Stenslie. Towards Telehaptic Performativity. Remote Encounters: Connecting bodies, collapsing spaces and temporal ubiquity in networked performance (2013).
12. H. Stone, J. Sidel, S. Oliver, A. Woolsey, and R.C. Singleton. Sensory Evaluation by Quantitative Descriptive Analysis. Food Technology (1974), 28. 24-34.
13. C. Wang, P. Cesar, and E. Geelhoed. An Invisible Gorilla: Is It a Matter of Focus of Attention? Proceedings of the Pacific-Rim Conference on Multimedia, (2013).

4 Virtual Microphone – Listening Experiments

Konrad Kowalczyk¹, Alexandra Craciun¹, Christian Dachmann¹, Nikolaus Färber¹

¹ Fraunhofer IIS, Erlangen, Germany

4.1 Introduction

In this section, a user-based experimental evaluation study of the *virtual microphone* (VM) technology developed at Fraunhofer IIS is presented. The virtual microphone technique aims to synthesize the signal of a non-existing (virtual) microphone placed arbitrarily in an acoustic space that sounds perceptually similar to the signal that would be recorded using a physical (real) microphone located in the same position. Such a VM signal is generated based on the spatial and acoustic information gathered by analysing the audio signals captured by at least two distantly placed distributed microphone arrays.

During the on-going development of this VM technique, many standard objective measures typically used in audio signal processing were applied to evaluate its performance, such as the signal-to-noise ratio, the direct-to-diffuse signal ratio, etc. However, since the main goal of the VM is to generate a signal that is perceptually similar to the signal of a real microphone placed in the same position, the subjective evaluation of the audio quality and perceptual similarity of such a virtually generated signal is of great interest and importance. The listening experiments presented here were conducted in lab-like conditions, thus the experiments are repeatable and the results can be used as a proof of concept for the proposed technology. All experiments were conducted according to the best practice of subjective audio evaluations.

This section is structured as follows. In Section 4.2 the purpose of performing three subsequent listening experiments is presented and short literature review of applied evaluation methods is provided in Section 4.3.1. The procedure of performing these experiments is described in detail in Section 4.3.2, followed by the description of the setup during recordings. The setup for playing signals to the listeners is provided in Section 4.3.3. The description of the speech signals used as stimuli and the categories of participants for the listening experiments are presented in Sections 4.3.4 and 4.3.5, respectively. Statistical analysis applied to the gathered results is presented in Section 4.3.6. The results together with a short discussion are provided in Section 4.4, followed by concluding remarks presented in Section 4.5.

4.2 Aim

In order to evaluate if the *virtual microphone* (VM) signals are perceptually close to the signals that would be recorded with physical microphones placed in the same locations, we performed three listening experiments to verify the following hypothesis:

- Experiment 1 – *Distance Perception*

Hypothesis: The listeners perceive a change in the distance from a sound source to different VM positions in a similar way as it would be perceived when using real microphones placed at the same distances as for the VM.

- Experiment 2 – *Angle Perception*

Hypothesis: The directions of sound sources captured using virtual stereo microphone recording techniques are perceived by listeners in a similar way as they would be perceived when recorded using real stereo microphones.

- Experiment 3 – *Perceptual Similarity of the Spatial Image*

Hypothesis: The spatial image captured using stereo VMs sounds perceptually similar to the spatial image recorded using real stereo microphones located at the same position.

In this listening test, we aim to find out if the listener feels as if he was in the same position in the sound scene for both virtual and real stereo recordings, focusing on the perceived spatial image and not on signal coloration. Perceptual similarity concerns also the possible quality degradation or perceived artefacts in *virtually* generated sound.

4.3 Method

4.3.1 Literature review

This section presents a short literature review and motivation for the selection of the listening test procedure presented in the next section. There exist several methods for subjective evaluation of audio quality, typically applied to evaluate the quality of audio codecs [1]. For example, the ABX test relies on identifying a reference signal X as A or B, where one signal is a hidden reference and the other is a coded version of X. Depending on the p-value, the result is that there is (or is not) a perceivable difference between samples. The BS.1116 test relies on assigning grades on a continuous scale to A and B, in comparison to a known reference. These tests are suitable for finding out if a given codec results in a significantly different quality than the original signal. In a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test [2], [3], a number of test samples (also called conditions), a hidden reference and at least one anchor are rated on a scale between 0 and 100. The purpose of using an anchor is to obtain a closer-to-absolute scale, such that small artefacts are not judged as very bad quality. This test is very well suited for comparing the audio quality of several codes.

Considering the spatial image of the sound scene [4], the human perception of the position of sound sources in a recorded sound scene is based on the perceived angles from which the source signals arrive and the distances to them. Also the amount of reverberation plays an important role, in particular the ratio between the direct and the reverberant sound. If the distance to the source is small, the direct-to-reverberation ratio is high and the loudness level is high as well. For large distances to the source, there is much less direct sound, the reverberant sound dominates, and the loudness level is significantly lower [4]. Estimating the distance to the source based on audio signals only is a very difficult task. It has been shown in several studies [5], [6] that listeners tend to overestimate the distance to the source for the distance below 1 meter and underestimate it for larger distances. The discrepancy between the physical and the perceived distances can exceed 2 times the actual physical distances for 0.5 or 5 meters [5], [6]. On one hand, the human perception of directions of arrival of sound sources is quite accurate for sound sources in front of a listener [7]. On the other hand, angular hearing resolution decreases for sound sources coming more from the side directions.

Since we aim to test the spatial image perception and audio quality of a VM signal, a standard MUSHRA test is selected as a testing method in experiment 3. An additional advantage of this approach is that it offers the opportunity to compare a number of signal processing blocks (such as localization algorithms, single [8] or multichannel signal extraction [9] methods) and learn which one of them leads to better perceptual quality of the VM signals, in comparison with a reference physical-microphone recording. However, a MUSHRA test yields just one result per tested condition (VM version) and does not allow us to find out if important elements of the human perception of spatial sound are correctly recreated in VM signals. Two crucial attributes of the spatial image when moving a VM in a sound scene are the distance and angles from which the sound sources arrive. Therefore, separate tests are defined to verify if the distance and angle are correctly preserved in the VM signals. In order to take limitations of the human perception into account in the experiment [5], [7], we compare the perceptual results of real and virtual microphone signals, and do not directly compare them with the absolute (physical) distance and angle values.

4.3.2 Procedure

The virtual microphone experiment consisted of three subsequent listening tests, organized in two sessions of 30 minutes duration each. In the first session, the distance and angle perception were evaluated using both real and virtual stereo microphone recordings. These two attributes were selected as separate tests as they are crucial for the human perception of spatial image when listening to the sound scene recorded at a given position. Since the goal of the virtual microphone (VM) technology is to generate the signal of a non-existing (virtual) microphone placed arbitrarily in the sound scene, the distance and angle from which the source signals are impinging on such a VM are very important to our perception of the listener position in the sound scene. On the other hand, we are also interested in verifying how close perceptually the VM and real recordings are for microphones placed at the same location. Therefore, in the third experiment, which was run in the second session organised a few days later, we aimed to investigate how close perceptually the real and virtually generated microphone signals are, thus verifying to what extent achieving the main goal of the VM technique is possible.

In all experiments, the stereo signals were recorded using a virtual and real stereo microphone setup referred to as XY-stereo technique. XY stereophony is a standard stereo recording technique [4] in which two microphones are placed at the same position with a 90° angle between their look directions and the stereo effect is achieved by differences in sound intensity between the two microphones.

The real stereo signals were recorded in room with low reverberation at Fraunhofer IIS in predefined positions. In addition, the signals captured using two distributed (that is positioned distantly) microphone arrays were recorded. The latter signals were then processed off-line using the VM software and the VM stereo signals were generated. Such real and generated stereo audio files were played back over two loudspeakers to the participants in a listening room. The details of the recording and listening conditions and the setups used in each test are all described and presented in illustrative figures in Section 4.3.3. For the source signals, four different speaker signals (2 male and 2 female) in English and German were selected (see Section 4.3.4 for details). Finally, 28 and 30 participants respectively took part in the two listening test sessions, including both expert listeners and unexperienced listeners (see Section 4.3.5 for a detailed description).

Experiment 1 – Distance Perception

In this experiment, we aimed to evaluate the distance perception for stereo recordings using real microphones at three defined distances and compare them to the distance perception for stereo VM recordings at the same positions. The source was always positioned in front of a listener and the stereo recordings were performed at the following distances: 0.75m, 1.5m, and 3m. Four different male and female speakers were used, as described in Section 4.3.4.

By comparing the statistics of the results (such as means, variances, and confidence intervals) obtained independently for both real and virtual stereo microphones, one could deduce how well the distance is perceived in general and how close the distance perception of the proposed VM method is compared to real recordings. Note that this way the limitations of the human ear for distance perception were also taken into account since we did not aim to obtain a perfect match between the perceived distance values provided by the listeners and real listener-source distances.

The experiment consisted of 11 sets, each containing 3 recordings at different distances. The listeners were asked to assess how far the source was from their position on a continuous scale from 0 to 4 meters. The first set was a training set, using different source-listener distances than those used in the actual test. After the assessment of the first 3 distances recorded using real microphones, the listener was presented with the true distance values. It should be noted that the order of presentation of different speakers in 10 subsequent

distance sets was randomized automatically, mixing both real and virtual recordings. Similarly, the three presented distances were ordered randomly in each set.

The part of the listener instructions related to the purpose and tasks during this experiment was: *In this experiment, you will listen to a sound source which is always located in front of you. The distance between you and the source will however be changing. The task of the experiment is to assess how far away the source is located, where the distance to the source can vary between 0 and 4m. The experiment consists of 11 sets, each of them containing 3 scenarios corresponding to different source distances.*

The software for running the distance test was written with the purpose of performing the VM listening experiment described in this report. The listener could switch between three scenarios (different distances) and was asked to indicate the perceived distance to the source from the range of 0 to 4 m using a slider. The listener could switch as many times as required between the three scenarios, start, stop, and loop over a specific part of the playback. After three distance values were set and accepted as final answers, the listener moved to the next set of three distance scenarios, and could neither go back to the previous set nor change previously selected distances. A screen shot of the tool for the distance assessment is presented in Figure 4.1.

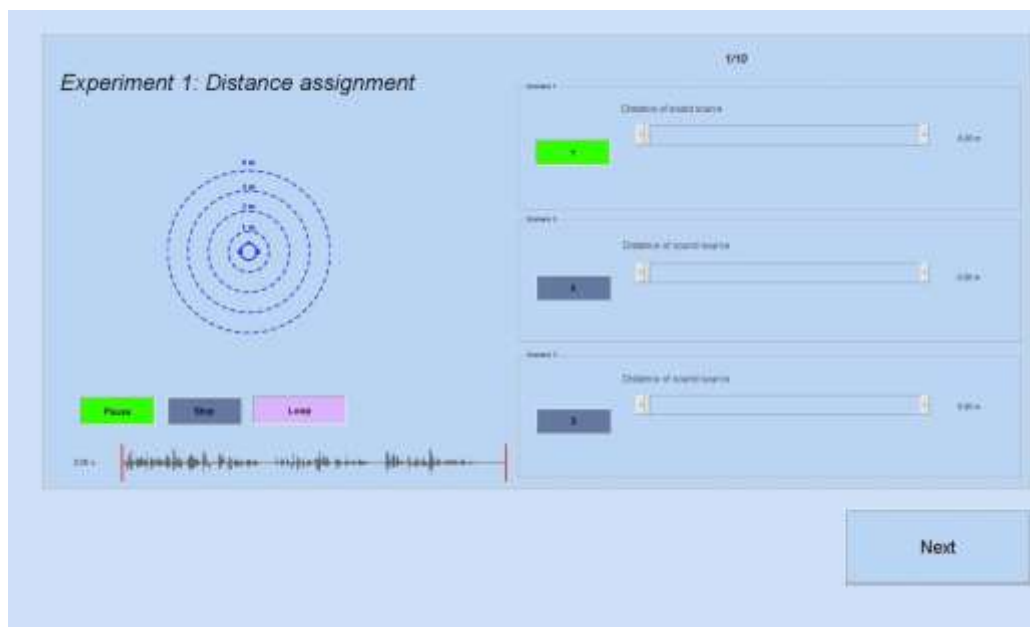


Figure 4.1: GUI for assessing the distance in Experiment 1.

Experiment 2 – Angle Perception

The angle perception test aimed to verify if the angle from which the source signal arrives is perceived similarly in the spatial stereo recording using VM and real stereo microphones. For this purpose, the real and virtual stereo microphones were positioned in 3 different positions in a room, and a sound source was speaking from one of two defined source positions. This allowed us to capture the speaker signal from 5 different directions, namely from -57° , -37° , 0° , 37° , and 57° , respectively. The details of the measurement setup are described in Section 4.3.3. Different speaker signals were again used, as described in Section 4.3.4. Similarly to the previous experiment, the statistics of the results (such as means, variances, and confidence intervals) obtained independently for both real and virtual stereo microphones were compared. Although in general, one could also relate these perceived angular values to the true angles of arrival.

The experiment consisted of 22 sets, each containing one recording. The listeners were asked to estimate the perceived angle from which the sound source was speaking using a continuous scale. The listeners were

additionally separated from the loudspeakers with a curtain such that no visual cues related to the room setup could be used while assessing the angle of arrival of the sound source. To make it easier for the listener to point in the selected direction, the angular plot was always provided on the left side of the GUI, where the angle the source was recorded at was remapped to the corresponding angles of a stereo setup. The first two sets were training sets, where after providing the selected answer; the listener was presented with the true angle value. It should be noted that the order of presentation of different speakers, coming from different angles in 20 subsequent sets was randomized automatically, mixing both real and virtual recordings.

The software for running the angle test was written with the purpose of performing the VM listening experiment described in this report. The listener could start, stop, and loop over a specific part of the playback. After the angle value was selected and accepted as a final answer, the listener moved to the next scenario and could neither go back to the previous set nor change previously selected distances. A screen shot of the tool for assessment is presented in Figure 4.2, the arrow in the left plot in the GUI indicating the selected angle.

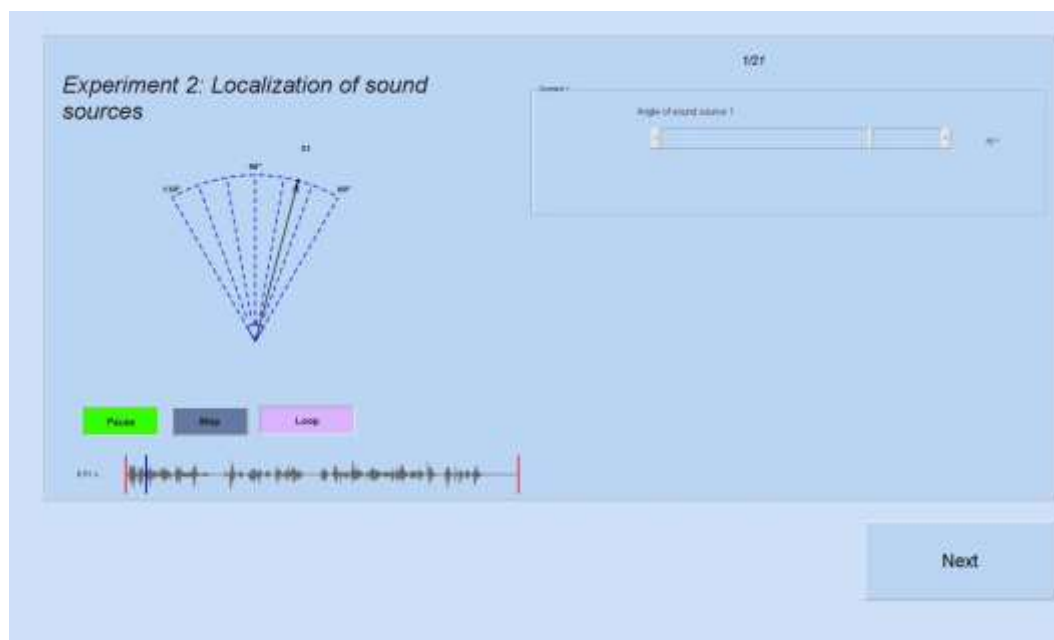


Figure 4.2: GUI for source angle perception in Experiment 2.

Experiment 3 - Perceptual Similarity of the Spatial Image

The third experiment aimed to verify how close perceptually the signals of the virtually generated and real microphone signals placed at the same location in the sound scene are. In particular, the goal was to find out if the listener perceives the sound scene as if he was placed in exactly the same position in both real and virtual recordings. In addition, sound degradation and audio artefacts caused by the VM processing should also be possible to capture by Experiment 3. Therefore, we decided to adopt a subjective audio quality evaluation method based on the Recommendation ITU-R BS.1534-1 [2], [3] called Multiple Stimuli with Hidden Reference and Anchor (MUSHRA), which is a widely used method for subjective evaluation of the quality of audio codecs.

In the MUSHRA test, the listener is presented with a reference signal (in our case a stereo recording using real microphones), to which a condition, for example a virtual microphone recording, can be compared. Since this test was originally designed for comparison of the performance of different codecs [2], [3], it offers an option

to use more conditions (in our case different versions of virtual microphone recordings) in a test set, thus allowing for the comparison of the performance of various VM signal processing blocks and parameter settings. This way, not only it is possible to verify how close perceptually the VM technique was in comparison to the real recording from the same position, but also allows to assess which signal processing blocks lead to the best performance in terms of the perceived spatial image and audio quality. In addition to these tested VM conditions, two more conditions need to be presented to the listener in a MUSHRA test: a hidden reference and an anchor. Using a continuous grating scale, the listener assigns a score in the range of 0 to 100 to each of the tested conditions, which corresponds to a 5-point grading as depicted in Figure 4.3. A hidden reference has to be identified by the listener by setting its score to 100 and a lower anchor should be defined such that the full grating scale is used. However, it is not required that listeners grade the lower anchor as 0.

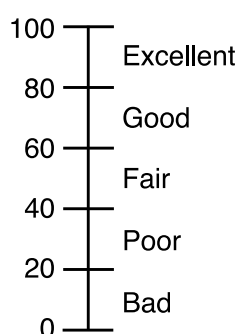


Figure 4.3: Continuous rating scale for MUSHRA.

The MUSHRA listening test for spatial image perception was run in a separate (second) session of around 30 minute duration, which was performed a few days after the first session. The experiment consisted of 5 sets, each with 6 different conditions to be evaluated. These 5 sets (also referred to as scenarios) included both single speaker recordings and recordings of the sound scene with two simultaneously speaking sources, a detailed description of the measurement setups being described in Section 4.3.3. As for the assessment of the VM method, 4 different conditions were tested: VM with multichannel signal extraction (with optimum and non-optimum parameter settings) [9], VM with single-channel signal extraction [8] and a real-time capable implementation of VM that uses a simplified localization algorithm. Note that the earlier two VM methods require off-line processing. For an open reference, a recording using real microphones placed at the same position as for 4 virtual microphones was used. As a lower anchor, a recording using real microphones from a large distance to the sources was selected, such that the spatial image (especially the directions of the sources) was hardly perceived due to the lack of strong direct sound. Note that typically in MUSHRA tests the lower anchor is obtained by distorting the reference signal [2]. However, since the aim of this experiment was to evaluate the spatial image, no additional distortions were introduced to the anchor signal.

The part of the listener instructions related to the purpose and tasks during this experiment was: *In this experiment you will listen to various sound scenes, where the spatial image of the sound sources composing the scene will be changing. The task is to assess how similar the spatial image of each sound scene is with respect to the spatial image of a reference sound scene. You will be listening to 5 test sets in total, each set containing 6 conditions to be graded.*

Please grade the overall spatial image of each sound scene by comparing with the reference sound scene. To do this, please consider the following characteristics:

- *the position of the source(s) in the sound scene,*
- *the broadness of the source(s),*
- *the ratio between direct and reverberant sound.*

Please do not judge based on coloration or timbre.

For running the experiments, a MUSHRA software written at Fraunhofer IIS was used, a screen shot of which is presented in Figure 4.4. The listener could start, stop, and loop over a selected part of the playback. In order to conveniently compare different test conditions, the listener could instantaneously switch between any of the 6 conditions and the reference. Once the rating was assigned to all test conditions, the listener moved to the next set (scenario), and could neither go back to the previous set nor changed previously selected MUSRHA scores.

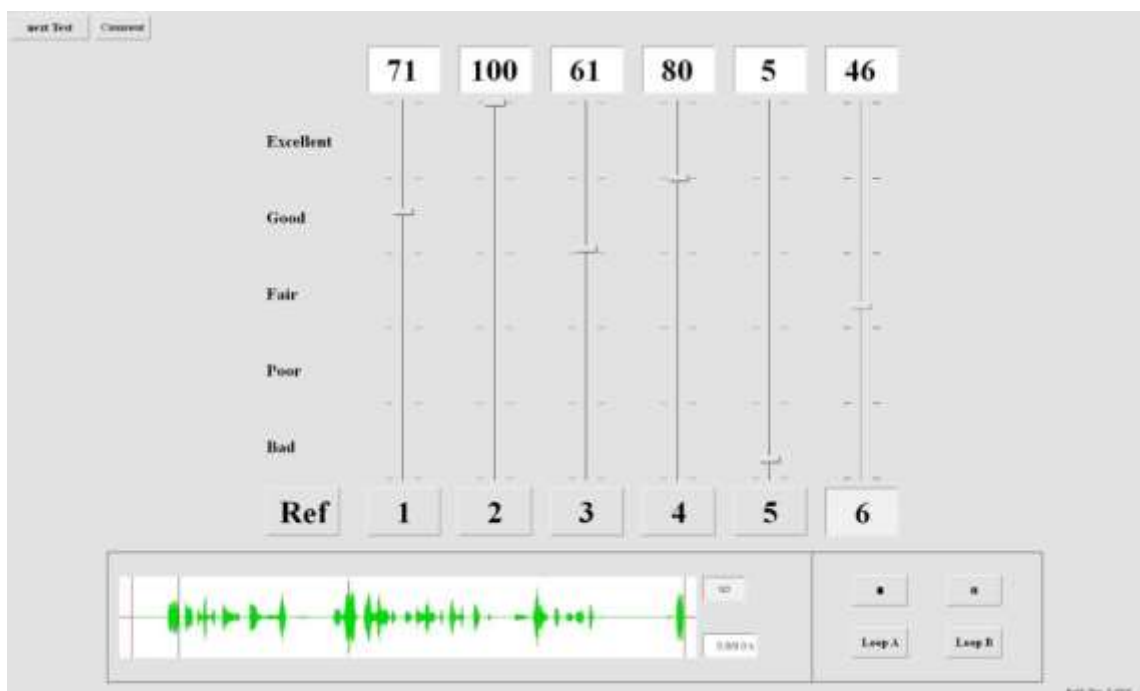


Figure 4.4: GUI used in the MUSHRA test.

4.3.3 Experimental Setup

Sound recording

Room: The listening experiment was based on signal recordings performed in one of the rooms at Fraunhofer IIS as shown in Figure 4.5. The Mozart room is large 9.3m x 7.5m x 4.2m yet despite its size it is characterised by a low reverberation time of 0.35 seconds. The floor is covered by a thin carpet and the walls and ceiling are covered by a partially absorbing material. This room was selected for this experiment because of its low reverberation characteristics (note that high reverberation is the main challenge in capturing distant audio signals) and because of a large empty area in the middle of the room for placing the equipment required to record signals for the experiments.



Figure 4.5: Mozart room at Fraunhofer IIS in which microphone signals were recorded.

Equipment: The speech signals were reproduced in a room using standard audio Genelec loudspeakers shown in Figure 4.6(a). To acquire the input audio signals needed to generate the VM signals, two microphone arrays, each equipped with four cardioid microphones AKG CK31 with inter-microphone spacing of 0.02 m were used, as depicted in Figure 4.6(b). In addition, microphone pairs were placed in dedicated positions in the room such that real stereo signals were recorded for those positions. These real stereo signals were used as reference for the comparison with the virtually generated stereo signals for the same locations. The same AKG microphone type was used for the stereo XY recordings.



(a)



(b)

Figure 4.6: (a) Genelec 8030APM loudspeaker used as a sound source and (b) Microphone array with AGK CK31 microphones.

Setup for all experiments: Two uniform linear microphone arrays were used to capture sound, as shown in Figures 4.7-4.9. The right array was rotated by $+20^\circ$ and the left array was rotated by -20° (that is both arrays were rotated towards each other). The distance between the arrays was 1.5 meters, and their positions were kept constant across all experiments. The number of speakers (sound sources) varied between 1 and 2 across experiments, resembling the typical real-life scenarios. All sound sources were positioned 1.5 meters in front of these two arrays, either in front of one of the arrays or in front of a point close to the centre between both

arrays. In each experiment, we defined 3-4 positions, where real and virtual microphones were placed. In the following, we describe the setup for each experiment in detail.

Setup for Experiment 1

The experimental setup for the distance perception experiment is shown in Figure 4.7. A sound source was located 1.5 meters in front of a point lying in between of two arrays, at a distance of 0.55 meter from Array 1 and 0.95 meter from Array 2. The sound scene was recorded using both microphone arrays and additionally stereo microphones (a pair of cardioid microphones) that were placed at 3 selected positions in a room. Stereo microphones were always placed in front of a source at distances 0.75, 1.5, and 3m away from the sound source.

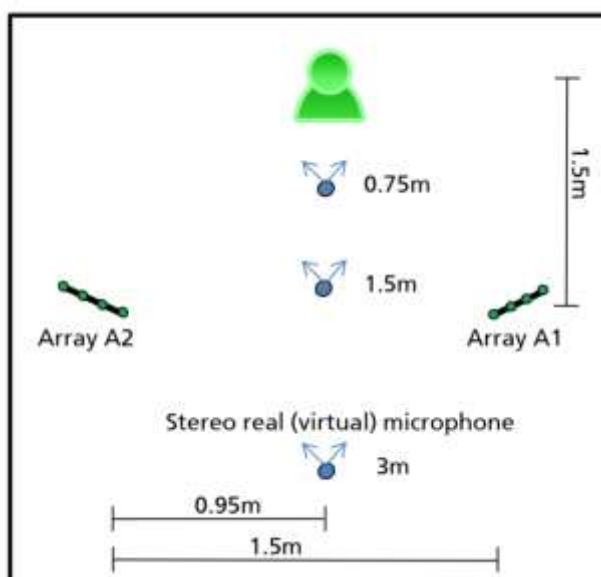


Figure 4.7: The measurement setup used in Experiment 1 for distance perception.

Setup for Experiment 2

The experimental setup for the angle perception experiment is shown in Figure 4.8. Two loudspeakers were placed, one in front of each array, at a distance of 1.5 meter from the closest array. However, only one source signal was played at a time, either speaker 1 or speaker 2, thus the other speaker was always silent. The sound scene was recorded using two microphone arrays and additional stereo microphones (a pair of cardioid microphones) placed at 3 defined positions in the room. Stereo recordings were taken along a line at 1 meter distance parallel to the line linking the sources. This setup allowed us to record signals impinging on the stereo microphones from a total of 5 different angles: -57° , -37° , 0° , 37° , and 57° .

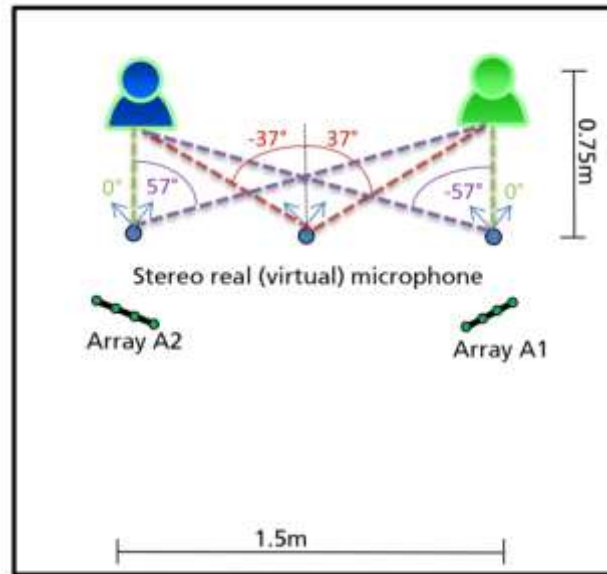


Figure 4.8: The measurement setup used in Experiment 2 for angle perception.

Setup for Experiment 3

The experimental setup for the MUSHRA test is depicted in Figure 4.9. Two loudspeakers were placed, one in front of each array, at a distance of 1.5 meter from the closest array. In this experiment, either 1 or 2 speakers were active at a time, and the sound was recorded using stereo microphones as done in Experiment 2 (along a line 1 meter away from the source line) and additionally at a position in front of one source at a distance of 0.75 meter away from the source. To record the lower anchor required for the MUSHRA test, the stereo microphones were rotated by 180° and they were positioned at a large distance to the sources: 4.1 meter behind the arrays and moved to the right direction by 3.17 meter.

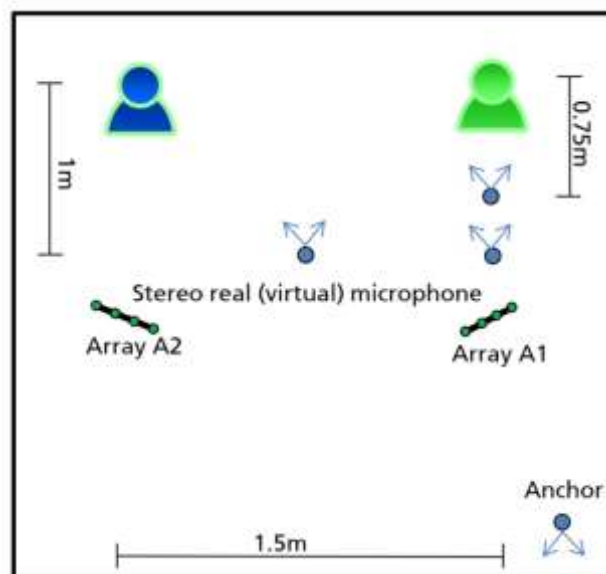


Figure 4.9: The source and microphone setup used for the MUSHRA test.

Listening room

Test participants were listening to the stereo recordings in one of the reproduction studios at Fraunhofer IIS, as shown in Figure 4.10. The room was acoustically treated such that its influence on the reproduced sound is minimised. Stereo audio files were played in a room through 2 loudspeakers located 2 meters apart from each other, where each was located 2 meters away from the listener position (sweet spot). Both loudspeakers were facing the listener and a curtain was put between the listener and the loudspeakers such that the perceived spatial image was not affected by visual cues related to the loudspeakers positions. This is typically done in listening experiments where spatial cues are of interest. The listener was sitting in front of a computer, which was used to control the interface of the test software and collect participant responses.



Figure 4.10: Sound reproduction in a listening room.

Angle remapping in Experiment 2:

One important aspect for the angle perception test was to map the angles from the real recording angles to angles that could be reproduced using a stereo loudspeaker system arranged in an equilateral triangle with 2 meter distances between the loudspeakers and between each loudspeaker and a source. The angular spread of the recordings was less than 120° (-57° to 57°) and the loudspeaker setup was able to reproduce signals with a spread of 60° (-30° to 30°). This angle mapping problem is illustrated in Figure 4.11, where the top figure presents 5 angles from which the source signal was arriving at the stereo microphones during the sound scene recordings, whereas the bottom figure presents the loudspeaker arrangement with reference to the listener and 5 angles of arrival for true source positions remapped to this particular loudspeaker arrangement. Listeners were asked to indicate the perceived angle of arrival for an angular spread of 60° , therefore participants results were later remapped to the angular spread of the original recording (120°) according to [10] before the data analysis was performed.

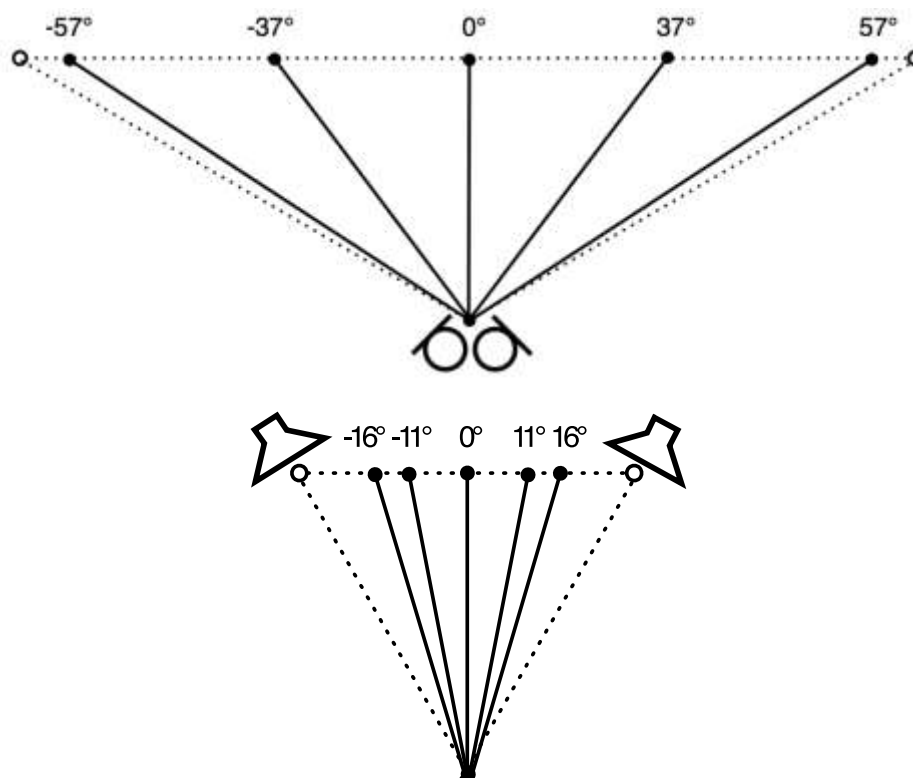


Figure 4.11: The angles of arrival of 5 source positions in the stereo recordings (top) and the loudspeaker setup with 5 remapped angles of arrival (bottom).

4.3.4 Stimulus (Speech Signals)

In order to ensure that the signals recorded at the distributed microphone arrays and at the real stereo microphone positions always correspond to the same source signal and that the measurement is repeatable, speech signals were played back via loudspeakers placed at the source positions, as explained in the previous subsection. These mono speech signals, often referred to as stimuli, were created as excerpts of speech of 7 to 12 seconds duration, as required by formal listening tests such as MUSHRA. These short audio sequences were all created from four longer speech signals taken from the EBU SQAM database [11]:

- Female speech English (Track number 49),
- Male speech English (Track number 50),
- Female speech German (Track number 53),
- Male speech German (Track number 54).

4.3.5 Participants

The experiments were run in 2 sessions. In the first 30 minute session, the distance and angle experiments were performed, in which the total of 30 participants took part. In the second 30 minute session, the MUSHRA test was performed by 28 participants. All participants were employees of Fraunhofer IIS or postgraduate students at International Audio Laboratories, a chair at the University of Erlangen-Nuremberg co-financed by Fraunhofer IIS.

The listeners could also be separated into 2 groups: expert and non-expert listeners. Expert listeners are the experienced listeners that often take part in listening tests, mainly related with audio codec evaluations. Non-

experts were considered students of International Audio Laboratories Erlangen and co-workers at Fraunhofer, who have not yet done a listening test or do not have much experience with listening tests. 14 experts took part in the distance and angle tests, and 13 experts participated in the MUSHRA test. The age range for expert listeners was 26 – 54 and for the non-expert listeners 22 – 38.

4.3.6 Statistical Analysis

Distance and angle perception experiments

To compare the distance and angle measurements for real and VM stereo recordings with each other, the means and the confidence intervals *CI* were calculated using (1) and (2), respectively. These analysis results are presented in Section 4.4. The calculation was done for all three distances and five directions of arrival according to:

$$\bar{u}_{jk} = \frac{1}{N} \sum_{i=1}^N u_{ijk} \quad (1)$$

where u_{ijk} is the score of participant i for the test condition j (for example different stereo recording positions) and speech stimulus k , where N is the number of listeners. The confidence intervals associated with the means were calculated based on the standard deviation and the sample size using:

$$[\bar{u}_{jk} - \delta_{jk}, \bar{u}_{jk} + \delta_{jk}] \quad (2)$$

$$\delta_{jk} = t_{0.05} \frac{S_{jk}}{\sqrt{N}} \quad (3)$$

where S_{jk} denotes the standard deviation which was calculated as:

$$S_{jk} = \sqrt{\frac{\sum_{i=1}^N (\bar{u}_{jk} - u_{ijk})^2}{(N-1)}} \quad (4).$$

The t -value for a significance level of 95% ($t_{0.05}$) was taken from a t -value table for a normal distribution, and for the sample size of Experiments 1 and 2 they corresponded to where $t = 2.045$. The confidence intervals were calculated this way based on the central limit theorem, despite the fact that the results were actually not normally distributed. Thanks to the central limit theorem, when the distribution of the population has a finite variance, the distribution of the arithmetic mean of random samples can be considered approximately normal in case of a sufficiently large sample size, as was the case for our experiments.

Additionally, a non-parametric test called the Wilcoxon-signed-rank-test was performed to obtain the p -values and the significance levels between real and VM results. These results were used to verify whether the medians were significantly different or not.

In preparation for this test, the histograms were first plotted, and it was verified that the compared distributions have the same shape. The following hypotheses were made:

H_0 : There is no difference between the medians of the results for a real and virtual microphone.

H_1 : There is a difference between the medians of the results for a real and virtual microphone.

Results were obtained using MATLAB and are shown in tables in Section 4.4. The smaller the p -value, the more can H_0 be rejected. For small values of p ($p \leq 0.05$), one could conclude that there is a significant difference between the medians of the compared recordings.

MUSHRA test

Before analysing the results of the MUSHRA test, the performance of each listener was checked and a post-screening was performed. The outlier results, which were found according to the method presented in the ITU-Recommendation [2], were taken out in order to ensure a more realistic assessment. As advised in the ITU-Recommendation [2], a listener was rejected if the hidden reference was rated below 90 in more than 15% of the items. A listener was also rejected if more than 25% of the items were rated higher than 1.5 times the third quartile Q_3 or lower than 1.5 times the first quartile Q_1 of all results. These first and third quartiles were calculated as follows:

$$Q_1(u) = \begin{cases} \text{median}\left(u_1, \dots, u_{\frac{n+1}{2}}\right), & \text{if } n \text{ odd} \\ \text{median}\left(u_1, \dots, u_{\frac{n}{2}}\right), & \text{if } n \text{ even} \end{cases} \quad (5)$$

$$Q_3(u) = \begin{cases} \text{median}\left(u_{\frac{n+1}{2}}, \dots, u_n\right), & \text{if } n \text{ odd} \\ \text{median}\left(u_{\frac{n}{2}}, \dots, u_n\right), & \text{if } n \text{ even} \end{cases} \quad (6)$$

Having removed the outliers, the mean and the confidence intervals were calculated for each test condition (4 versions of VM, anchor, and hidden reference). Not only the results for all listeners and audio stimuli were calculated and presented in figures, but also the results for one or two simultaneously speaking sources, as well as the results for expert and non-expert listeners were calculated and plotted.

In the MUSHRA plots, the results were shown as a combination of boxplots and means plus confidence interval for each test condition (version of the VM, anchor, and hidden reference). Boxplots required the calculation of medians, the inter-quartile ranges (*IQR*), and the lower and upper whiskers, which were computed using the following formulas:

$$Q_2(u) = \text{median}(u) = \begin{cases} \frac{u_{\frac{n+1}{2}}}{2} & \text{if } n \text{ odd} \\ \frac{1}{2\left(u_{\frac{n}{2}} + u_{\frac{n}{2}+1}\right)} & \text{if } n \text{ even} \end{cases} \quad (7)$$

$$\text{lower whisker} = Q_1 - 1.5 * IQR \quad (8)$$

$$\text{upper whisker} = Q_3 + 1.5 * IQR \quad (9)$$

$$IQR = Q_3 - Q_1 \quad (10)$$

4.4 Results and Discussion

4.4.1 Experiment 1 – Distance Perception

Figure 4.12 shows the listener evaluation in the distance perception experiment, where the results for the real (physical) microphones are depicted with blue colour and the VM results are depicted with red colour. For each of three tested distances, the mean is marked as a small circle and the confidence intervals are denoted by the ends of the whiskers. A black dashed line denotes perceived distances corresponding exactly to the physical distances (based on true measurement positions).

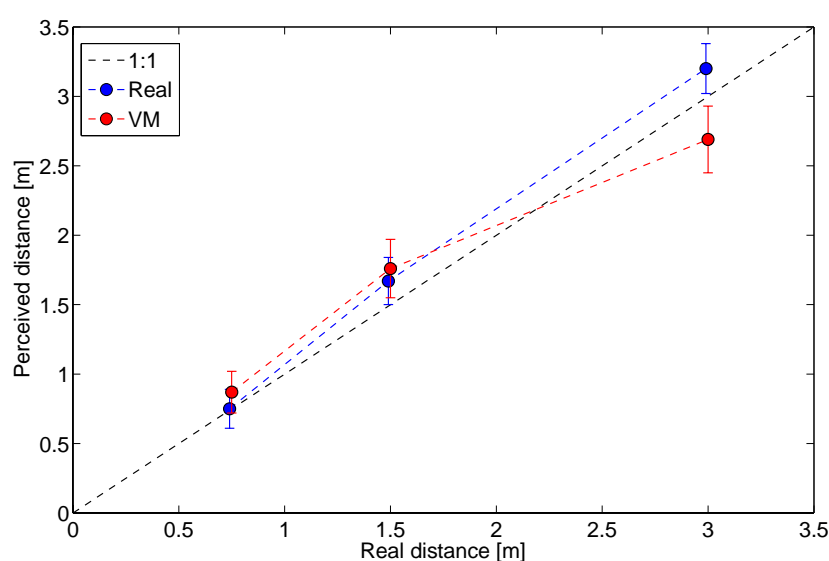


Figure 4.12: Mean values and confidence intervals in distance perception experiment.

As can be observed, from Figure 4.12, the means and confidence intervals for both real and virtual microphones are matching quite well and they relate closely enough to the real distances measured during the recordings. This result is a bit surprising as from the literature we would have expected stronger deviations of the perceived distances from the true ones, irrespectively if real or virtual microphones were used [6].

The distance perceived from VM recordings is overestimated at short distances and it is underestimated at long distances. In order to be able to verify if the perceived differences in listener-source distances are significant, the p -values of the Wilcoxon-test are listed in Table 2. As can be concluded from the computed p -values, there is a significant difference for the closest and furthest distances, whereas there is no significant difference for the middle tested distance.

Table 1: Results of the Wilcoxon test for the distance test.

| Distance | p-value (Wilcoxon) | significant difference |
|----------|--------------------|------------------------|
| 0.75m | 0.03 | yes |
| 1.5m | 0.3825 | no |
| 3.m | $3.896 * 10^{-6}$ | yes |

4.4.2 Experiment 2 – Angle Perception

Figure 4.13 depicts the results of listener evaluation for the experiment on the perceived angle of arrival, where the results for the real (physical) microphones are depicted with blue colour and the VM results are depicted with red colour. For each of the five tested angles, the mean is marked as a small circle and the confidence intervals are denoted by the ends of the whiskers. A black dashed line denotes perceived angles of arrival corresponding exactly to the true angles measured during the experiments.

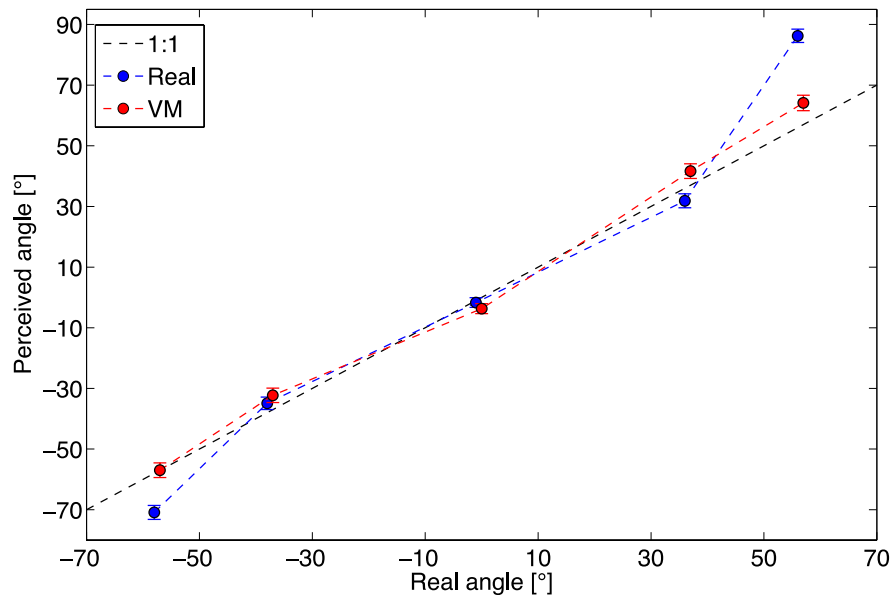


Figure 4.13: Mean and confidence intervals for angle test

As can be seen, the means and confidence intervals for both real and virtual microphones are again matching quite well. They are also close to the angles corresponding to the true source positions during the recordings. The largest deviations can be observed for -57° and 57° for the case where real microphones were used, which can be explained by imprecisions in positioning the real stereo microphones for those two cases during the recordings. Even if these positions were correct, it is possible that the two microphones used in the XY stereophony were slightly rotated, which could explain the angular shift in these results.

In order to be able to check if the differences between the angles of arrival for the real and virtual recordings are significant, p -values of the Wilcoxon-test were computed and listed in Table 2. As can be concluded from the computed p -values that do not exceed 0.05, there is no significant difference for 0° and -37° angles, whereas for larger angles the difference is significant.

Table 2: Results of the Wilcoxon test for the angle test.

| Angle | p -value (Wilcoxon) | significant difference |
|-------------|-----------------------|------------------------|
| -57° | 0.0011 | yes |
| -37° | 0.253 | no |
| 0° | 0.3413 | no |
| 37° | 0.0172 | yes |
| 57° | $1.468 * 10^{-6}$ | yes |

4.4.3 Experiment 3 – MUSHRA test

In this section, the results of the MUSHRA test for perceptual similarity between the real and virtual recordings are presented. These results indicate how close perceptually the VM signals are in comparison with the recordings using real microphones placed at the same location. As described in Section 4.3.2, four different conditions were tested for the VM processing, as well as a hidden reference and a lower anchor. The labels for these six conditions are presented in Table 3.

Table 3: Legend for the MUSHRA result figures

| | |
|-----|---|
| VM1 | VM multichannel (after [9]) |
| VM2 | VM single-channel enhancement (after [8]) |
| VM3 | VM real-time implementation (simplified localization algorithm) |
| VM4 | VM multichannel with optimum parameter selection for this particular room |
| AN | Anchor (real recording from a large distance to the source(s)) |
| HR | Hidden reference (real recording) |

As recommended in [2],[3], the boxplots for each tested condition are shown in the MUSHRA result figures presented below. Such a boxplot consists of a box ranging from the 1st to the 3rd quartile, with the band inside the box indicating the median (2nd quartile). The upper and lower whiskers starting from the top and bottom end of the box are also shown, both ending at the associated quartile +/- 1.5 times the inter-quartile-range (IQR). The calculation of these values was described in Section 4.3.6. Results larger than this range are marked as outliers (denoted as red plus). From such box plots, the centrality, spread and symmetry of the data can be observed.

For completeness of presentation, the means and confidence intervals are also provided next (left) to the boxplot for each condition.

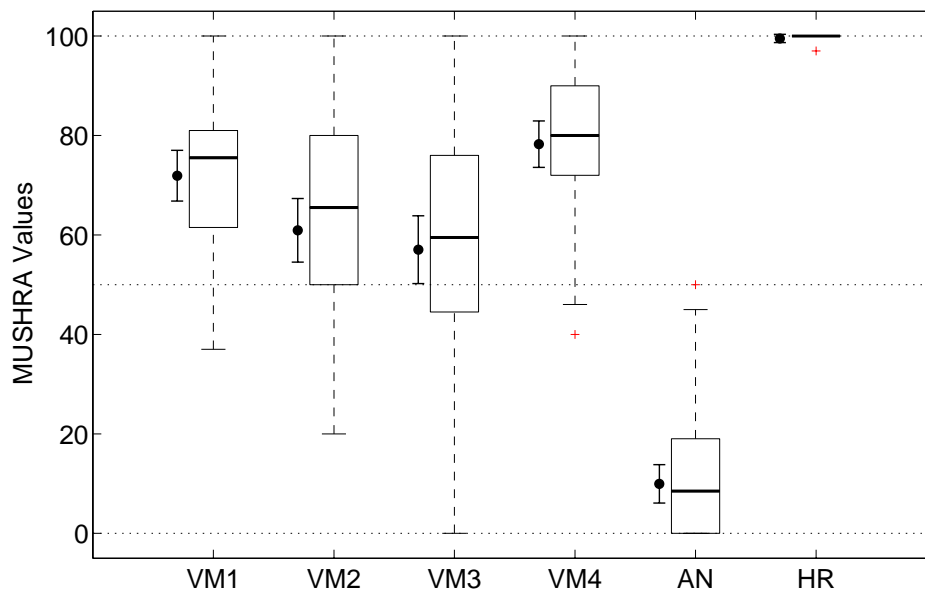


Figure 4.14: MUSHRA test results for all participants and all signal sets.

Figure 4.14 shows the results for all listeners and scenarios that were used in the test. It can be seen that in general the reference was found correctly (rating of 100) and the lower anchor was typically rated quite low. The spread of the results for VM2 and VM3 is large, whereas VM1 and VM4 were generally rated more consistently. Looking at the medians, we notice that VM3 was rated low, while VM1 and VM4 were rated best out of the 4 tested VM conditions, obtaining good and even almost excellent grade on a quality scale. Inspecting the overlap of the confidence intervals for different conditions, the quality of the spatial image in VM1 and VM4 can be considered significantly different than that of VM2 and VM3, and closer to the reference (real) recording.

However, simply inspecting the overlap of the confidence intervals is not enough to formally claim that various VM conditions are significantly different. For this reason, we conducted Wilcoxon-test for pairs of tested conditions, which were created by setting all conditions in the descending order of their median values and then taking a pair of neighbouring conditions. Such results for MUSHRA test shown in Figure 4.14 are listed in Table 4. As can be concluded, the reference recording was significantly different from all VM recordings, and differences between VM conditions were significant apart from the difference between VM2 and VM3. All VM conditions were also significantly different from the anchor recording

Table 4: Results of the Wilcoxon test for MUSHRA results shown in Figure 4.14

| pair | p-value (Wilcoxon) | significant difference |
|-----------|---------------------|------------------------|
| HR / VM4 | $1.539 * 10^{-44}$ | Yes |
| VM4 / VM1 | 0.0035 | Yes |
| VM1 / VM2 | $2.6095 * 10^{-5}$ | Yes |
| VM2 / VM3 | 0.0894 | No |
| VM3 / AN | $1.3936 * 10^{-37}$ | Yes |

We can also plot the results for 1 and 2 source scenarios, thus investigating if the performance of different VMs varies depending on the complexity of the recorded scenario. The MUSHRA test results for 1 source are presented in Figure 4.15, whereas the results for 2 simultaneously active sources are shown in Figure 4.16. In general, the quality of the spatial image for VM1 and VM4 was rated higher than for VM2 and VM3.

However, the performance of VM versions within those pairs varies depending on the complexity of the scenario. While VM4 outperforms VM1 in 1 source scenarios, VM1 seems to perform better in 2 source scenarios. Similarly, VM2 performs better than VM3 in 1 source scenarios but the opposite holds for 2 source scenarios. The p -values for the 1 and 2 source scenarios are presented in Tables 5 and 6, respectively. In 1-source scenarios, the differences between multichannel and single-channel VM versions, that is VM4 and VM1, and VM2 and VM3, respectively, can be considered significant. Furthermore, one cannot state that the difference between VM1 and VM2 is significant for the scenarios with 1 source only. On the other hand, in the 2 source scenarios, for the abovementioned pairs, VM4 and VM1, and VM2 and VM3, respectively, one also cannot conclude that the differences are significant. However, the difference between multichannel (VM4 and VM1) and single-channel processing (VM2 and VM3) is definitely significant.

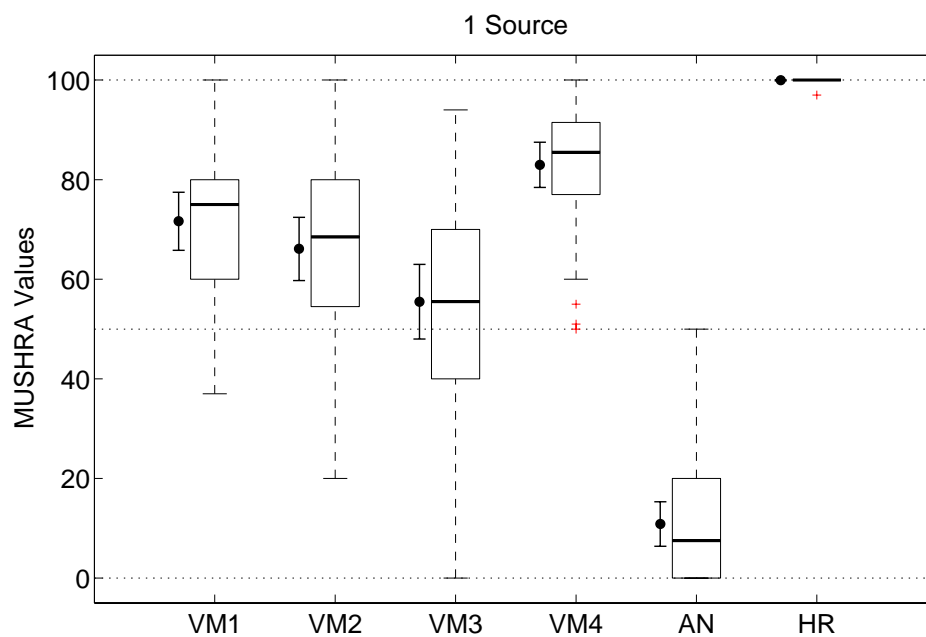


Figure 4.15 MUSHRA test results for all participants and one source scenarios.

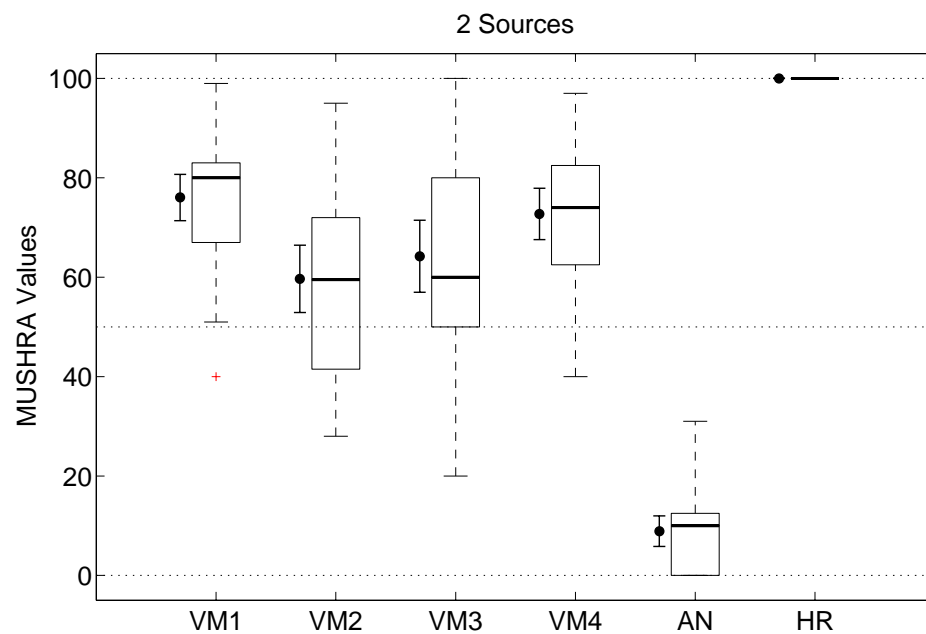


Figure 4.16: MUSHRA test results for all participants and two source scenarios.

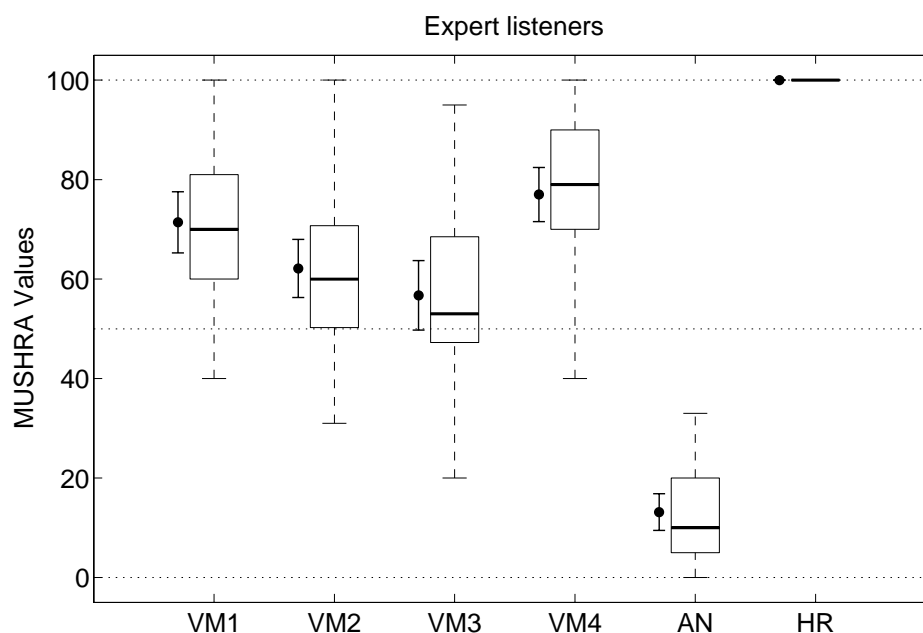
Table 5: Results of the Wilcoxon test for MUSHRA (1 Source)

| pair | p-value (Wilcoxon) | significant difference |
|-----------|---------------------|------------------------|
| HR / VM4 | $1.5385 * 10^{-26}$ | yes |
| VM4 / VM1 | $8.6579 * 10^{-6}$ | yes |
| VM1 / VM2 | 0.0830 | no |
| VM2 / VM3 | 0.0010 | yes |
| VM3 / AN | $1.1126 * 10^{-21}$ | yes |

Table 6: Results of the Wilcoxon test for MUSHRA (2 Sources)

| pair | p-value (Wilcoxon) | significant difference |
|-----------|---------------------|------------------------|
| HR / VM1 | $1.7543 * 10^{-19}$ | yes |
| VM1 / VM4 | 0.2170 | no |
| VM4 / VM2 | $4.1141 * 10^{-4}$ | yes |
| VM2 / VM3 | 0.1736 | no |
| VM3 / AN | $4.1342 * 10^{-17}$ | yes |

Finally, Figure 4.17 shows the MUSHRA results for expert listeners only, from which it can be seen that the same general trends are preserved as in the case when all listeners were considered (see Figure 4.14). However, expert listeners are much more consistent when rating the quality of the spatial image, shown by a smaller spread of results. For completeness, the p -values for different VM conditions as rated by expert listeners are presented in Tables 7.


Figure 4.17: MUSHRA test results for expert listeners and all scenarios.



Based on expert listener ratings, it can be concluded that the results for VM1 and VM4 are again significantly different from (better than) VM2 and VM3, thus offering a spatial image and audio quality that is significantly closer to the image in real stereo recordings than the real-time capable VM3 version.

Table 7: Results of the Wilcoxon test for MUSHRA (Expert listeners)

| pair | p-value (Wilcoxon) | significant difference |
|--------------|--------------------|------------------------|
| Ref / VM4 | $1.591 * 10^{-21}$ | yes |
| VM4 / VM1 | 0.0829 | no |
| VM1 / VM2 | 0.0046 | yes |
| VM2 / VM3 | 0.0794 | no |
| VM3 / anchor | $2.776 * 10^{-18}$ | yes |

4.5 Conclusions

The conducted listening experiments, whose results were presented in this report, confirmed that the *virtual microphone* technology is generally capable of generating a signal of a non-existing microphone placed in a given position inside an acoustic space, which is close enough to the signal from a real microphone placed in the same position. The positions of the sound sources in such recordings are quite well matched with the source positions that would be recorded using real microphones placed in the same positions as the positions of VMs.

For distance and angle perception, the VM technique yields performance which does not differ significantly from the real recordings, especially for middle-range distances and angles of arrival of the sound sources. However, for more extreme values, such as large angles and large and small distances, the perceived difference in VM recordings can be considered significant. When comparing the audio quality and the spatial image of the VM with real recordings, it can be concluded that off-line multichannel processing leads to an at least good (almost excellent) match with real recordings on a standard audio quality scale. However, the VM processing that can be run in real-time on a powerful personal computer is rated much lower than previously mentioned VM versions, which is due to the use of a simplified yet less accurate localization algorithm required for real-time processing. As could be expected, a worse VM performance is obtained when moving a VM through an acoustic space with less accurately localized sound sources.

References

- [1] S. Bech and N. Zacharov, *Perceptual audio evaluation: theory, method and application*, Wiley, Chichester, UK (2006)
- [2] Recommendation ITU-R BS.1534-1, *Method for the subjective assessment of intermediate quality level of coding systems* (2003)
- [3] F. Nagel, T. Sporer, P. Sedlmeier, *Towards a Statistically Well-Founded Evaluation of Listening Tests*, Audio Engineering Convention, Paper 8146, London, UK (2010)
- [4] J. Blauert, *Spatial hearing*, Rev. ed. Cambridge, MA: The MIT Press (1995)

-
- [5] H.-Y. Kim, Y. Suzuki, S. Takane, and T. Sone, Control of auditory distance perception based on the auditory parallax model, *Applied Acoustics*, vol. 62, pp. 245-270 (2001)
- [6] P. Zahorik, Auditory display of sound source distance, *Proc. Int. conf. on Auditory Display*, Kyoto, Japan (2002)
- [7] J. Blauert, Sound localization in the median plane, *Acustica*, vol. 22, pp. 205-213 (1969)
- [8] O. Thiergart, G. Del Galdo, M. Taseska, E. Habets, Geometry based spatial sound acquisition using distributed microphone arrays, *IEEE Trans. Audio, Speech, Language Process.*, vol. 21 (12), pp. 2583-2594 (2013)
- [9] K. Kowalczyk, O. Thiergart, A. Craciun, E. Habets, Sound acquisition in noisy and reverberant environments, *Proc. IEEE Workshop Appl. of Sign. Process. to Audio and Acoustics (WASPAA)*, New Paltz, NY (2013)
- [10] E. Sengpiel, Stereo-Lautsprecherlokalisierung eines 120°-Klangkörpers @ONLINE, 2013. <http://www.sengpielaudio.com/Stereo-LautlokEines120.pdf>
- [11] Sound Quality Assessment Material recordings for subjective tests @ONLINE, 2013. URL <http://tech.ebu.ch/publications/sqamcd>



5 A QoE Testbed for Socially-Aware Video-Mediated Group Communication

| | | | |
|--|--|--|--|
| Marwin Schmitt | Simon Gunkel | Pablo Cesar | Peter Hughes |
| CWI: Centrum Wiskunde & Informatika | CWI: Centrum Wiskunde & Informatika | CWI: Centrum Wiskunde & Informatika | BT Research & Technology |
| The Netherlands | The Netherlands | The Netherlands | United Kingdom |
| schmitt@cw.nl | gunkel@cw.nl | p.s.cesar@cw.nl | peter.j.hughes@bt.com |

ABSTRACT

Video-Mediated group communication is filtering into everyday use, as commercial products enable people to connect with friends and relatives. Current solutions provide basic support, so that communication can happen, but do they enable conversations? This paper argues that the purpose and the context of the conversation are influential factors that are rarely taken into consideration. The aim should be on the development of underlying mechanisms that can seamlessly palliate the effects of networking variances (e.g., delays) and optimize media and connection for every single participant. In particular, our interest is on how to improve remote multi-party gatherings by dynamically adjusting network and communication parameters, depending on the ongoing conversation. If we are to provide a software component that can, in real-time, monitor the Quality of Experience (QoE), we would have to carry out extensive experiments under different varying (but controllable) conditions. Unfortunately, there are no tools available that provide us the required fine-grained level of control. This paper reports on our efforts implementing such a testbed. It provides the experiment conductor with the possibility of modifying and monitoring network and media conditions in real-time.

5.1 INTRODUCTION

As video-mediated group communication gradually finds its way into our everyday life, we need to build systems that support our needs whether we are just casually catching up with family overseas or watching the latest game of our favorite sports team with far away friends. Current systems adapt the media to the network conditions [1] or change the layout to portray the loudest location (Google+ Hangout).

To build socially-aware systems, we need a better understanding on how asymmetric network conditions, group activities and different roles affect the individual and the overall QoE (e.g., how delay on a single participant affects the overall QoE? Should we provide more bandwidth to active participants over passive participants? Should we use more of the available bandwidth for frame-rate or resolution?)

Previous research investigated dyadic conversations [2], high-end business-oriented solutions [3] and the implication of de-synchronization when watching videos remotely [4]. But they only provide us with an starting point. The ITU is starting to look into the direction of QoE assessment of multiparty tele-meetings [5]. Still there are no recommendations available and the current knowledge is not sufficient to build systems, which can act upon the influencing factors of QoE.

Such knowledge is obtained through extensive user trials under diverse, but controlled, conditions. Unfortunately, none of the publicly available solutions provide the flexibility and level of control, which is required for extensively investigate the influence of network and media parameters on the QoE. We investigated how these experiments can be done with Google's Hangout but we ran into several problems. The control and manipulation of the technical aspects are only indirectly possible through simulating network

conditions. If we are to investigate asymmetric network conditions this requires an extensive infrastructure. Monitoring the experiment sessions becomes also quite problematic. In standard video-conferencing software the experiment conductor cannot be hidden, which influences the trial. Solutions for recording the media streams, in the original and degraded version, are either accompanied by quality reduction, which does not allow reasoning about the original perceived quality or require expensive specialized hardware.

This paper tries to fill a current gap: the lack of an adequate testbed for controlled experiments, which allows obtaining conclusive results regarding QoE in video-mediated group communication. It describes our solution the Video-Mediated Communication Testbed (VMC-TB).

5.2 QoE in Video Mediated Communication

In order to evaluate the QoE in video mediated communication (VMC) we have to address two issues: *what factors* do we look at and *how do we measure* them.

5.2.1 Factors of QoE in VMC

The model we use is based on the framework by Geerts et al. [6] and the one by Wu et al. [7]. We applied them to VMC and consider in this paper only aspects for controlled experiments. This model, visualized in Figure 2, considers the different factors that will influence QoE in video-mediated communication. QoE, as a cognitive process, is located in the center. The distance of the individual factors to the center denotes how strongly they influence the current experience. The model has three dimensions: System, User and Context.

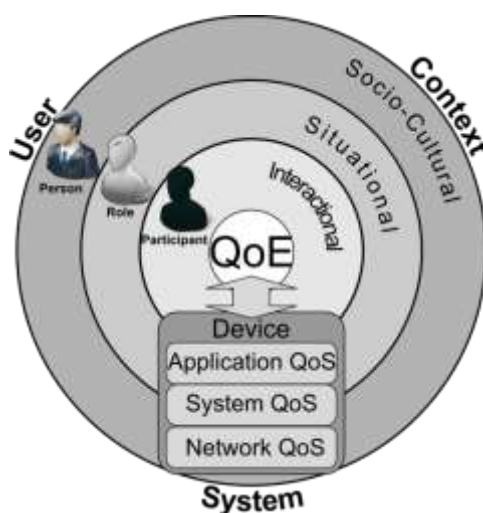


Figure 2 Influence Factor Model for QoE in VMC

The **System** dimension represents the technical aspects. This dimension is modeled as a stack. The lowest layer contains the network aspects, which influence the system layer, which in turn influence the application layer (the user perceptible aspect). At this layer is where QoS and QoE interface, as the user interacts with the system. Real-Time QoS monitoring usually traces network aspects like packet delay, jitter and bandwidth, but users only judge the end result like video artifacts and audio quality. The impact of the network level on the application level is shaped by numerous factors like protocol and codecs. The parameters of QoS impacting QoE in VMC, based on the model for distributed interactive systems by Wu et al. [7], are shown in Figure 3.

QoS stack is embedded in the device and accessed through the system UI, which also directly interacts with the user, in this paper we are focusing on dynamic aspects influenced by the network.

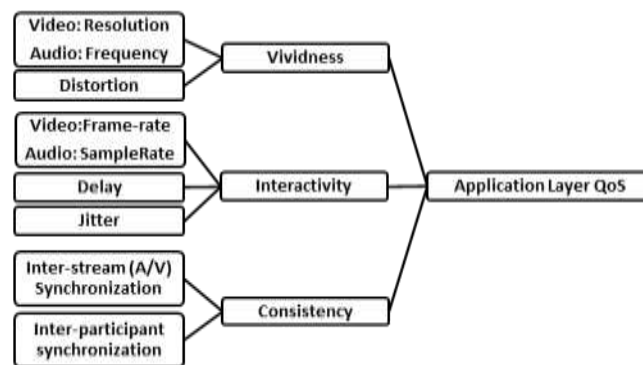


Figure 3 QoS Parameters interfacing QoE (adapted from [7])

The **user** dimension distinguishes between a person and the roles this person takes in video-mediated conversations [6]. On the participant level, we are concerned with experiences that are typical for a general user of a VMC system. On the role level, we want to detail experiences, based on the assumption, that a user adopts a specific role in a specific context. E.g. Given some delay, the experience from the moderator might be different from the one of other participants. Finally, at the person level we consider the individual experience this user has. The relation from more general to very individual experiences is illustrated in Figure 2 by the level of detail the user icon has.

The **context** is created by the interplay of the user with the situation. The user reacts to based on how she/he perceives the situation and in turn shapes it with her/his actions which makes the context inherently unstable [8]. The context can be classified based on three levels [6] [8]: interaction, situational and socio-cultural. The interactional context covers the interaction of the user with the system and the current task at hand. The interactional context is embedded in the situational context, which concerns the session in terms of activities and participants. The user has certain interests from which she/he forms, given the opportunities of the current situation, her/his current goals. The socio-cultural context, in which these situations take place, deals with aspects like societal conversation etiquette and habits. . The socio-cultural layer can be modeled as a reciprocal interaction between social norms and actions people evaluate, plan and carry out.

5.2.2 Measuring Methodologies

The impact of these factors is modeled through an adaption of the basic process from environmental psychology. In this process environmental influences form cognitive perceptions and lead to behavioral consequences [9]. It is still an ongoing research to determine which cognitive perceptions should be considered for QoE, but it is clear that we have to measure the behavioral consequences. The measurements are usually categorized into subjective and objective methods[7].

Subjective methods are self-reports from the user, giving insight about how the participant perceived the session from his or her own point of view.

Questionnaires are a common methodology for assessing impressions in a quantitative way. They allow processing the feedback with statistical methods.

Interviews are commonly used for gather qualitative feedback. They provide a descriptive view of the users experience and can provide more detailed insight about it.

Objective methods are based on externally observable and quantifiable behavioral changes.

Task Scores are metrics based on the task the participants have to perform in the experiment. Common metrics are task completion times or successful vs. unsuccessful attempts. Task scores are, in VMC, often only an indirect measure, since many scenarios do not have an inherent quantifiable task.

Speech Patterns are a more direct look at the ongoing interaction. For this a model of turn-taking [10] is applied which quantifies the ongoing interaction in terms of speaking times, length of turns and pauses,

simultaneous speech etc. Previous research has found influences on speech patterns from the previously detailed factors: mediating technology [11], roles [12] and context [13].

Physiological measurements are based on biological reactions (e.g. heart rate) which are correlated with experiences.

5.3 Video-Mediated Conversation-TestBed

In this section we present our developed testbed. We first give a quick overview of the system. Then we explain how the system enables us to investigate the factors highlighted in Section 2.1 and how the methodologies, detailed in Section 2.2, are integrated.

5.3.1 Video Client for Multiparty Conferencing

Our testbed consists of a Video-Client for Multiparty Conferencing, shown in Figure 4, an ObserverControl Client for the experiment conductor and a tool for analyzing experiment sessions (Figure 6).



Figure 4 VMC-TB Client

The clients are full-featured multiparty-video conferencing applications which are directly connected with each other. The system is designed so it runs in a controlled environment, at the moment we transmit data over User Datagram Protocol (UDP). We implemented the media processing pipelines of our testbed using GStreamer², a flexible, open-source toolkit with source-filter-sink based architecture. Figure 5 shows a simplified version of a sending and a receiving pipeline for the video stream. Besides the normal elements for capturing, encoding, and transmission, we added elements for monitoring and controlling the network and media parameters (see section 3.2.1). While GStreamer is implemented in C, we implemented the not-so performance critical components in the more lightweight programming language Python. This gives us a flexible platform, which is easily extensible and customizable.

The experiment conductor (using the ObserverControl Client), is usually not shown to the other participants, not to influence the trial, but can dynamically join the conversion, if necessary, to give feedback or additional instructions. Furthermore, the different steps of the experiment can be scripted based on the status of the system. E.g. to automatically show a questionnaire after a task is finished, set new conditions after all questionnaires are filled out and so forth. Each individual step is logged and available for data analysis.

5.3.2 Support for QoE Factors

In the following, we explain how the main QoE factors are supported in our testbed.

² www.gstreamer.org

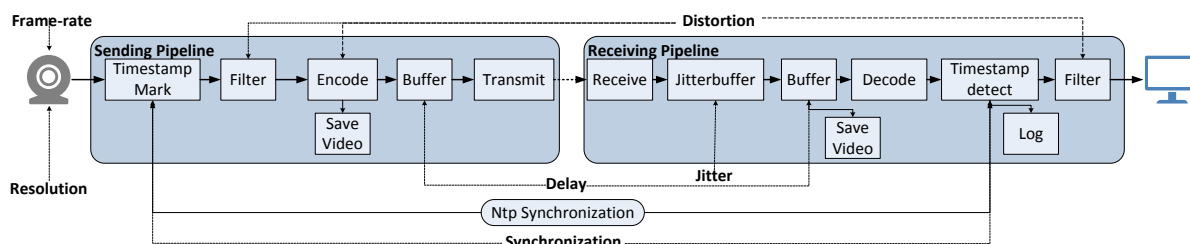


Figure 5 Video Pipelines in VMC-TB

5.3.2.1 QoS Parameters

In section 2.1 we identified some QoS parameters that will have an impact on the QoE. To investigate the effects of these parameters, we need to be able to control and monitor them. For monitoring the visual (for audio respectively acoustical) aspects, we record the transmitted streams on the sender and receiver side. To keep track of the temporal aspects, we synchronize the clocks of our clients via the Network Time Protocol (NTP) and log the delay of every frame. For this we directly insert a barcode into the video, which we crop-out at the receiving side before presenting the video to the user (compare Figure 4 and Figure 6). By directly inserting the timestamps into the video we measure the delay of the whole processing pipeline, instead of only of the network delay. For the complete “mouth-to-ear” delay we need to also consider the delay of capturing and rendering equipment, which can be assumed to be static and can be measured using external tools [14].

The parameters Resolution/Frequency and Frame-rate/Sample-rate can be manipulated directly at the corresponding capturing elements (with respect to the capabilities of the devices). For the other parameter, we use the following:

- **Distortion:** We can control distortion by inserting available filters from GStreamer (e.g. blur) or changing the codec settings. The easy extensible plugin architecture of GStreamer makes it easy to develop and integrate custom, more complex distortion patterns.
- **Delay:** The minimum delay our system achieves, in the ideal conditions of our local network, is in average 70ms with a 25ms standard deviation. We can add delay by increasing minimum amount of data hold in the buffers on the sending and the receiving side.
- **Jitter:** We keep the network delay constant by employing a jitterbuffer. We can add jitter by adjusting the buffer on the receiving side.
- **Interstream (Audio/Video) Synchronization:** We can achieve audio/video (de)synchronization by manipulating the delay buffers in audio and video streams separately.
- **Inter-participant Synchronization:** Since there is a separate pipeline for every participant we can achieve basic (de)synchronization by setting different delays for each participant. Since we have synchronized clocks and the capturing timestamps more complex synchronization algorithm can be built on top of this.

All parameters can be modified in real-time at a running system, which is required, since modern networks have typically fluctuating performances during one session [15].

5.3.2.2 User

We achieve the monitoring of each individual participant by recording the received and sent video at each client. This allows an investigation from the perspective of each user. The roles arise from the experiment design. Forcing users into formal roles can be achieved by assigning them to the participants in the experiment (e.g. discussion moderator) or creating scenarios with roles present (e.g. tele teaching with teacher and student). We support the experiment conductor in this task, by allowing him to label each participant with an individual a role. This label can be used to asymmetrically manipulate the

aforementioned parameters, execute specific behavior for the activity and is available as metadata for the data analysis after the experiment.

Further insights into the user's personal traits can be gained by integrating corresponding questions into the questionnaires before or after the experiments. Biases, attitudes and expectations with the experiment scenario and VMC systems in general should be considered.

5.3.2.3 Context

The interaction context is shaped by the user, when interacting with the system under a given task [6]. The User Interface of the VMC-TB Client is designed to support easy customization. It can be easily adapted for a specific experiment with the Glade GUI-Builder³.

For example, the client shown in Figure 4 is designed for an experiment that investigates the effect of delay in semi-structured group discussions with 5 participants. The client shows a small version of the own video in the upper left corner, a task specific pane below it and the other 4 participants in square layout. In the task pane, we implemented a shared view of the questions participants had to discuss and to select an answer. In this experiment we decided to make a static layout, which shows all participants in the same size, as we focused on delay, and wanted to keep the influence from layout as constant as possible. The UI is rather simple designed, based on the experience in UI design, that participants are more intimidated by prototypes which appear like completely finished software products and thus are more reluctant to give user feedback [16]. Further than the layout and task integration, the local context is shaped by the specific experiment design. The situational context is in controlled experiments always imposed ("participating in an experiment") and the socio-cultural context by inviting the appropriate participant to the study. Further insights into the socio-cultural background of the participants can be gained through questionnaires before or after the experiment session. Questionnaires to assess the socio-cultural background should thus investigate knowledge, experiences and plans of activities similar to the experiment activity and VMC in general.

³ glade.gnome.org

5.3.3 Assessing Feedback

Questionnaires

VMC-TB integrates functionality so questionnaires can be administered. The questionnaires can be easily defined over a simple document format and displayed to the participants throughout the experiment. The experiment conductor receives the results from the questionnaires directly after they have been filled out. The integration of the experiments into the testbed has many advantages. The questionnaire can be made an intrinsic part of the experiment, so completion of the questionnaire can trigger the next step of the experiment. Furthermore it is easy to dynamically integrate aspects of the session at hand, e.g. questions about specific participants or based on the completion of the task.

Task Scores

The integration of tasks makes it possible to make an automatized scoring of the outcome. In the example sketched in Section 3.2.3 the answers of the decision making task was transmitted to the observer during the experiment. Additionally, other more traditional measures such as completion time are easily obtained, as the testbed logs all experiment steps.

Speech Patterns

The transmitted media is saved on the sending and on the receiving side, this enables us to analyse of the conversation after the experiment from each each participants perspective. We created a tool for viewing an experiment session and analyzing speech patterns (see Figure 6). The tool shows for each participant a timeline with blocks when there is audio activity. These are usually speech (shown in light blue), and can be later tagged by the analyst by different categories, e.g. laughter shown in green or non-verbal utterance (e.g. “uh”, “hmm”) shown in orange.

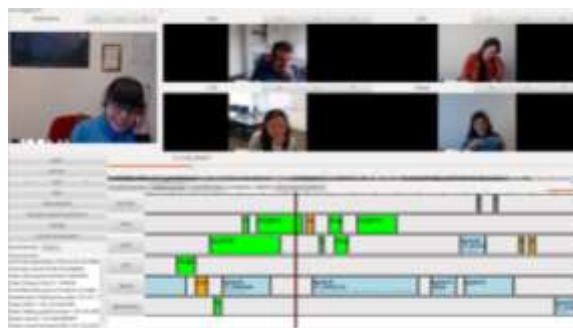


Figure 6 Speech Pattern Analysis Tool

Interviews and **physiological** measurements are not specially supported by our tool. Interviews are indirectly aided, as our system makes it easy for the experiment conductor to have a first look at the questionnaire data during or directly at the end of the trial. This makes it possible to focus in the debriefing on areas where additional information might be needed for the interpretation of the data.

5.4 DISCUSSION

We reported in this paper a testbed for conducting experimental research. We showed that the software permits us to control and directly manipulate a number of application level QoS parameters. The monitoring capabilities allow us, to analyze in detail the conducted experiments.

The tool allows us to consider different scenarios and activities. By integrating the task into our system, and monitor it, we can better understand what kind of interaction patterns arise and how they change under the influence of our QoS parameters. The detailed monitoring of each participant, allows us to investigate, whether specific effects are dependent on the role of the user.

With the analysis of speech pattern, we have a promising tool, which provides objective data that complements the subjective user feedback. The recording of the media streams in original and degraded quality, allow us to investigate whether current objective full-reference QoE metrics, correlate with the subjective feedback.

As a first step we started to investigate the effect of delay on semi-structured group discussion with 5 participants. We used a decision making task, similar to the survival scenario [5] and assigned one of the participants to be the moderator. As a starting point, we conducted a study with around 50 participants, setting symmetric conditions for all participants. In the next step, we will asymmetrically modify the delay conditions, to gain insight how this impacts the individual vs. the overall QoE of the group. This will give us insight what are appropriate inter-participant synchronization schemes for group discussions.

Further, we want to investigate the effect of the parameters video and audio quality and common VMC activities like sharing media. The testbed is a first step towards our final goal: detailing the effect from different factors (QoS, context, roles) on the individual and overall QoE in video-mediated group communication.



5.5 REFERENCES

- [1] L. De Cicco, S. Mascolo, and V. Palmisano, "Skype video responsiveness to bandwidth variations," in *Proc. of NOSSDAV'08*, 2008, pp. 81–86.
- [2] J. Tam, E. Carter, S. Kiesler, and J. Hodgins, "Video increases the perception of naturalness during remote interactions with latency," in *Proc. of CHI'12*, New York, NY, USA, 2012, pp. 2045–2050.
- [3] E. Geelhoed, A. Parker, D. J. Williams, and M. Groen, "Effects of Latency on Telepresence," HP labs technical report: HPL-2009-120 <http://www.hpl.hp.com/techreports/2009/HPL-2009-120.html>, 2009.
- [4] D. Geerts, I. Vaishnavi, R. Mekuria, O. Van Deventer, and P. Cesar, "Are we in sync?: synchronization requirements for watching online video together.," in *Proc. of CHI'11*, 2011, pp. 311–314.
- [5] P. 130. ITU-T RECOMMENDATION, "ITU-P.1301 - Subjective quality evaluation of audio and audiovisual multiparty telemeetings." 27-Feb-2013.
- [6] D. Geerts, K. De Moor, I. Ketyko, A. Jacobs, J. Van den Bergh, W. Joseph, L. Martens, and L. De Marez, "Linking an integrated framework with appropriate methods for measuring QoE," in *QoMEX'10*, 2010, pp. 158–163.
- [7] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang, "Quality of experience in distributed interactive multimedia environments: toward a theoretical framework," in *Proc. of ACM MM'09*, New York, NY, USA, 2009, pp. 481–490.
- [8] G. Mantovani, "Social Context in HCI: A New Framework for Mental Models, Cooperation, and Communication," *Cogn. Sci.*, vol. 20, no. 2, pp. 237–269, 1996.
- [9] A. M. and J. A. Russell, "An Approach to Environmental Psychology," Apr. 1980.
- [10] H. Sacks, E. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.
- [11] L. Ten Bosch, N. Oostdijk, and J. de Ruiter, "Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues," in *Text, Speech and Dialogue*, 2004, pp. 563–570.
- [12] W. Dong, B. Lepri, A. Cappelletti, A. S. Pentland, F. Pianesi, and M. Zancanaro, "Using the influence model to recognize functional roles in meetings," in *Proc. of ICMI'07*, New York, NY, USA, 2007, pp. 271–278.
- [13] M. Jochemsen, "The Effect of Context on Turn Taking in Human to Human Conversation," presented at the TSConIT'11, 2011.
- [14] J. Jansen and D. C. A. Bulterman, "User-Centric Video Delay Measurements," presented at the NOSSDAV, 2013.
- [15] I. T. S. Sector, "ITU-T Recommendation P. 880," *Contin. Eval. Time-Varying Speech Qual.*, 2004.
- [16] A. Black, "Visible planning on paper and on screen: The impact of working medium on decision-making by novice graphic designers," *Behav. Inf. Technol.*, vol. 9, no. 4, pp. 283–296, 1990.

6 Integration Vconnect into SAPO Campus: first experiment

Pedro Torres & Erik Geelhoed

6.1 Introduction

One of the aims of the current experiment is to study the interplay between social network and video-conference usage; in particular to evaluate the influence of a (seamless) integration of social network features with video-conference functionality on the user experience of video communication. Based on the view mode experiments it was decided that participants could choose whether to use the Hangout and/or Tiled view modes. Three different aspects are investigated: *Video-conference Starting*, *Browsing and Joining Conversations* and *User Interface*.

6.2 Method

6.2.1 Setup

Six rooms were connected using Vconnect's Home VClient integrated into the SAPO-Campus environment. The rooms were relatively big, classroom size, with four rooms on the ground floor and two rooms on a first floor. Participants used PC's running Windows, pre-prepared with the right plugins, drivers, Google Chrome, etc. on a wired Ethernet connection using standard quality headphones-microphones sets.

6.2.2 Participants

Twenty five participants, 8 females 17 males, mostly Communication and Art students in the age range 18-27 although there were two aged 31 and 45. All were registered regular users of the SAPO (DeCA) Campus

6.2.3 Experiment

Each experimental session lasted a total of two hours, including a 30 minute break. The participants were briefed for 20 minutes. They carried out three tasks lasting 15 minutes, followed by five minutes filling out a questionnaire. The sessions were concluded by a 20 minute group discussion.

The test session revolves around a debate (preferably with strongly opposing views) that participants will have on a topic chosen by them as a group. The sessions consist of three 10-minute stages aimed at exposing users to different features and covering different aspects and conditions of the videoconference integration with the social network. The participants could choose between three topics (rooms) for debate

- Social Networks: productivity enhancers or time wasters?
- Big Brother: functionality vs privacy
- A room called "Lounge", for free chat.

The whole room thing ended up not being very useful because, except in one case, everyone would convene in a single room and then discuss whatever they wanted, and most of the time conversations revolved around video-conferencing and the Vconnect/Sapo-Campus application they were using.

The three tasks consisted of group discussions around:

- Topic Selection
- Case Building
- Debate

During the first task, called *Topic Selection*, participants were encouraged to “hang-about”, i.e. to be present in the social network browsing and interacting with each other as well as inspecting the list of current conversations and entering and leaving video-conference rooms. This is a more passive and individual interaction, which can be interspersed with occasional video-chat. The idea was to expose participants to the social network usage and the specific use cases of conversation browsing and joining/leaving conversations. The specific task that users had at hand was to make up their mind about which topic they would like the debate but they were free to do whatever they wished during that period. After 10 minutes, the “conversation room” which would have the most people in it would be selected as the debate topic and used in the subsequent stage. In case of a draw, either the experimenter picks one randomly or the topmost of the tied topics in the room list will be selected.

In the next stage, called *Case Building*, participants were encouraged to pick a topic and “enter the corresponding room”. When users entered a room, they would be visible in the list of participants of that room. They were allowed to use other browser tabs to find relevant information or media. They were encouraged to write text and upload media to a SAPO Campus group as they prepare their case. This task had a time limit of 15 minutes from the time at least two people were in the room.

This team collaboration task would have the potential to expose users to two-, three-, and possibly four-way conversations within a chat room, conversation-related activity publishing, media sharing in SAPO Campus groups during video-conversation. It would result in an unstructured hangout since people would not have specific roles or constrained turn taking.

In the “*Debate*” stage, participants were given an opportunity to present their case in a six-way conversation. The aim was to have a slowish ping-pong (turn-taking) where, in turn, each side would put forward one side of the argument. The debate would last for as long as participants keep it going but there was a time limit of 15 minutes.

This stage exposed the participants to a six-way conversation with chat rooms and media sharing in SAPO Campus groups. It would be a more structured hangout since each side had a turn to present an argument and turn taking should be slower as, since they have pre-prepared material and thoughts, participants might tend to speak for longer periods, although they were allowed to interrupt each other.

6.2.4 Interviews, Questionnaires, Video footage and Automated Logging

Data collection consisted of group interviews, questionnaires and automated logging of button presses as well as conversational parameters. In addition all sessions were video recorded.

6.3 Results

Analysis is still in the initial stages, but will be carried out along similar lines as the Viewmode experiments.

Providing different topic-rooms ended up not being very useful because, except in one case, everyone would convene in a single room and then discuss whatever they wanted, and most of the time conversations revolved around video-conferencing and the Vconnect/Sapo-Campus application they were using.





However, from spontaneous comments made by the participants, the technology was well received.

"This is great! When can we start using it?"

"Sound was really good! Much better than Skype."

Participants had the option to choose different Viewmodes:

1. Hangout Style as part of SAPO Campus web page
2. Hangout style without being embedded in a webpage
3. Tiled as part of SAPO Campus web page
4. Tiled without being embedded in a webpage

| | |
|---|--|
|  |  |
| <p>Hangout Style as part of SAPO Campus web page</p> | <p>Hangout style without being embedded in a webpage</p> |
|  |  |
| <p>Tiled as part of SAPO Campus web page</p> | <p>Tiled without being embedded in a webpage</p> |

Different participants displayed different preferences for view modes, but overall they did not think using a different web-browser tab was a good idea.

[Sharing in a different tab?] *“NOT GOOD; we'd like to see it at the same time, e.g. below the video-conference.”*

Participants valued screen-sharing and would like to have a button on a SAPO Campus group which can trigger a conference in that group (groups-circles). This would be beneficial for ad-hoc group conferencing, but this would not involve all the people that were ever in “a room”, but only the people online at that point in time.

Most participants used point to point (free) Skype or Facetime with only a few using Google+ Hangout. Most of them being students, they would (much) rather use SAPO Campus video (Vconnect) conferencing in an educational context.

Automatic orchestration (triggered by voice activity) was received well but participants expressed the wish to manually override it. Only few participants preferred the Tiled view mode. This might be reflecting the predominance of males in the group of participants.

Participants explicitly asked: *“So, what's the maximum number of people we could get on this?”*

When answered that aim was around 30 people as a maximum, they thought this *“would be really cool”* but also acknowledged that the tiled or Hangout view modes would not be appropriate.

Participants were pleased with the audio quality. They would like to be able to set everyone's volume level individually.

Where “Subgrouping” is concerned, only one (out of the 25 participants) explicitly said that he would like to be able to have a “private” conversation with just a single person whilst still remaining part of the five-party video-conference. Participants were given the option to choose being in two “rooms” during a videoconferencing session but they would cluster in one single chat room.

6.4 Preliminary Conclusions

This mixture of laboratory and field trial, typical for Vconnect “coming out of the lab and into the field” proved to be a success. Participants were positive and indicated that they would like to use this kind of application as an extension of SAPO-Campus.

7 Relevant other main activities and connections

One concern that was expressed at the last review last January 2013 related to our level of involvement with the world of performers and performance; the target group in the performance use case. Of course it helps that both Falmouth and Goldsmiths Universities have Departments of Performing Arts and that experimental as well as developmental (CAVEs) work has involved Falmouth based performance lecturers and students in on-going informal discussions and student projects. In the GSR experiment the actors were closely involved making the GSR sensors “work”, by devising sections that required more or less active, i.e. physical, involvement by the audience.

In this section we report on a number of activities that anchor Vconnect’s research into distributed performance further into the performing arts community.

7.1 NESTA grant with Miracle Theatre

The National Endowment for Science, Technology and the Arts (NESTA) awarded a research grant to a small consortium consisting of Miracle Theatre (Cornwall based but touring nationally in the UK), Falmouth University, Dogbite Filmcrew (a Falmouth based Video, Film and Streaming company, e.g. they recently streamed a live performance of Jessie J to about 20.000 viewers) and the London based Golant Digital (entertainment) Media distributors.

The objectives of this 11 Months research project are:

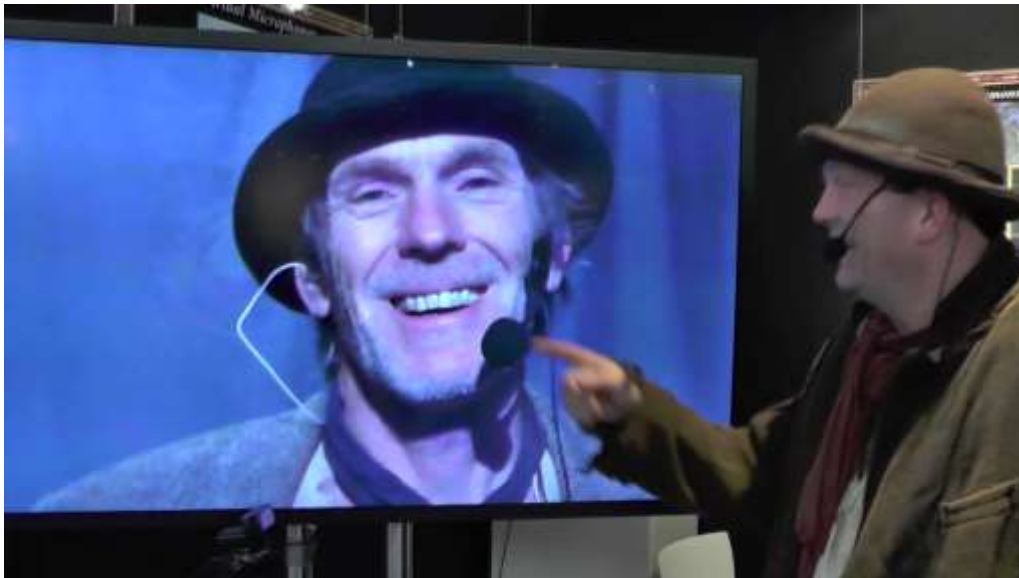
- To explore the interface between live and digital theatre performance
- To meet the challenges of reaching audiences in dispersed areas developing and testing new, affordable and exciting ways of distributing digital versions of live and recorded, productions amongst a rural network of venues and via on-line platforms.
- To experiment with a variety of low-budget digital methods of recording small-scale live performance to capture its particular qualities of intimacy, vitality, spontaneity and interaction and test the audience's experience of these.
- To test a sustainable financial model, establish appropriate pricing structures, marketing and rights agreements.

In September a live recording of Miracle Theatre performing *Waiting for Godot* (written by Samuel Becket) in the Performance Centre at Falmouth University was streamed to three remote venues (Manchester, Plymouth, St Agnes (in Cornwall)) and an audience evaluation was carried out in Falmouth and the three remote venues. We will report on this research within Vconnect as the research is relevant. In addition we will use sections of the recording to compare audience response to an edited (orchestrated) version with audience response to a static view of the performance.



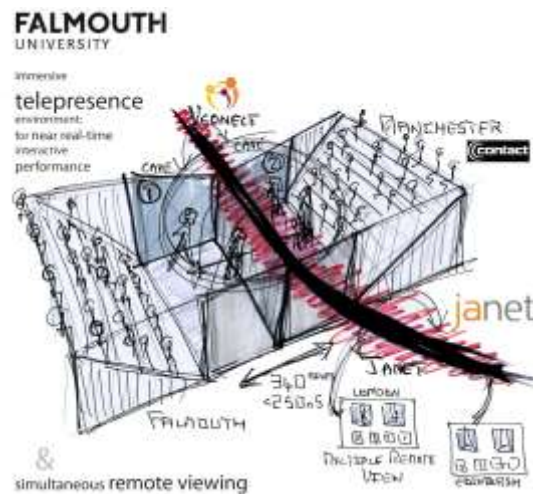
7.2 Vilnius Demo

Involvement of Miracle Theatre with Vconnect research was instrumental in making the recent demo at ICT2013 in Vilnius a success, as we demonstrated a remote performance of Samuel Becket's *Waiting for Godot*, where one actor resided in Falmouth and another in Vilnius. Using the Vconnect platform to deliver high resolution and low latency video images, they played live to an audience at ICT2013.



7.3 JANET connection and grant application

We presented our Vconnect remote performance work at "Remote Encounters" in Cardiff, April 2013 (<http://remote-encounters.tumblr.com/>), a conference on mediated communication and performance art. This resulted in establishing some on-going contacts with performance artists experimenting with this type of technology including developing close connections with a special interest group of the UK academic broadband JANET around delivering interactive mediated performance. Recently Falmouth University was successful in a grant application to the JANET's mediated performance research arm.



The proposal will use streaming solutions that have already been developed but in a new remote performance context, enabled by Janet's network. We will also test a larger scale deployment of additional remote viewers, supporting access to the performances beyond the two physical locations where performances occur, enabled by Janet's ability to support larger numbers of simultaneous streams across their network. We will explore how we can make use of Vconnect's Studio Client.

7.4 Fascinate Conference – Falmouth University, August 2013

Moreover we organised a well-attended conference at Falmouth University in this area where artist and technologists came together. <http://www.fascinateconference.com/>



The conference featured a showcase with 24 performances and 30 installations attended by around 350 people. The conference featured four keynote speeches, 17 paper presentations and seven workshop, as well as plenty of opportunities to socialise and network.

8 Conclusions

This second year has been highly productive and has seen a high level of integration between the technical team and Vconnect's user research. We conclude that the Vconnect system is now coming out of the lab and being applied in real life settings, both in the social and performance use cases. In addition there has been interest in the research community as well as commercial interest in our work and its potential applications.

9 Planning

For the next year, the last year of Vconnect, the planning is well defined.

9.1 Performance Use Case:

Late January: Lab experiment - Comparing Edited Vs. Non-Edited Video

This concerns a laboratory experiment exploring the value of orchestration, or to be more precise, the value of home-viewing an edited video registration of a live theatre performance in a comparison to a single camera wide shot. Using a within subjects design balanced for order, participants in individual sessions will watch footage (five minute clips) on a laptop of *Waiting for Godot* in an edited (orchestrated) version and as a static camera view.

Although this will be a relatively small scale experiment, there might be some interesting outcomes. Anecdotal evidence suggests that watching a streamed live theatre performance (e.g. a National Theatre Live production) registered using four cameras with a live audience present is different from watching a film. In addition when watching a live theatre performance an audience member sits in an allocated chair and has a static view of a play. In addition to the two conditions in this experiment we will also try to gauge in which respects a streamed (and edited) version of a live performance is different from watching a performance live, or watching a film version in a cinema.

22nd February: Disklavier – Falmouth, Royal Academy for Music (London), Yamaha & Goldsmiths

Audience Evaluation and Vconnect supported Audio-Visual Connection



Disklavier is the brand name for a group of piano-related products made by Yamaha Corporation.

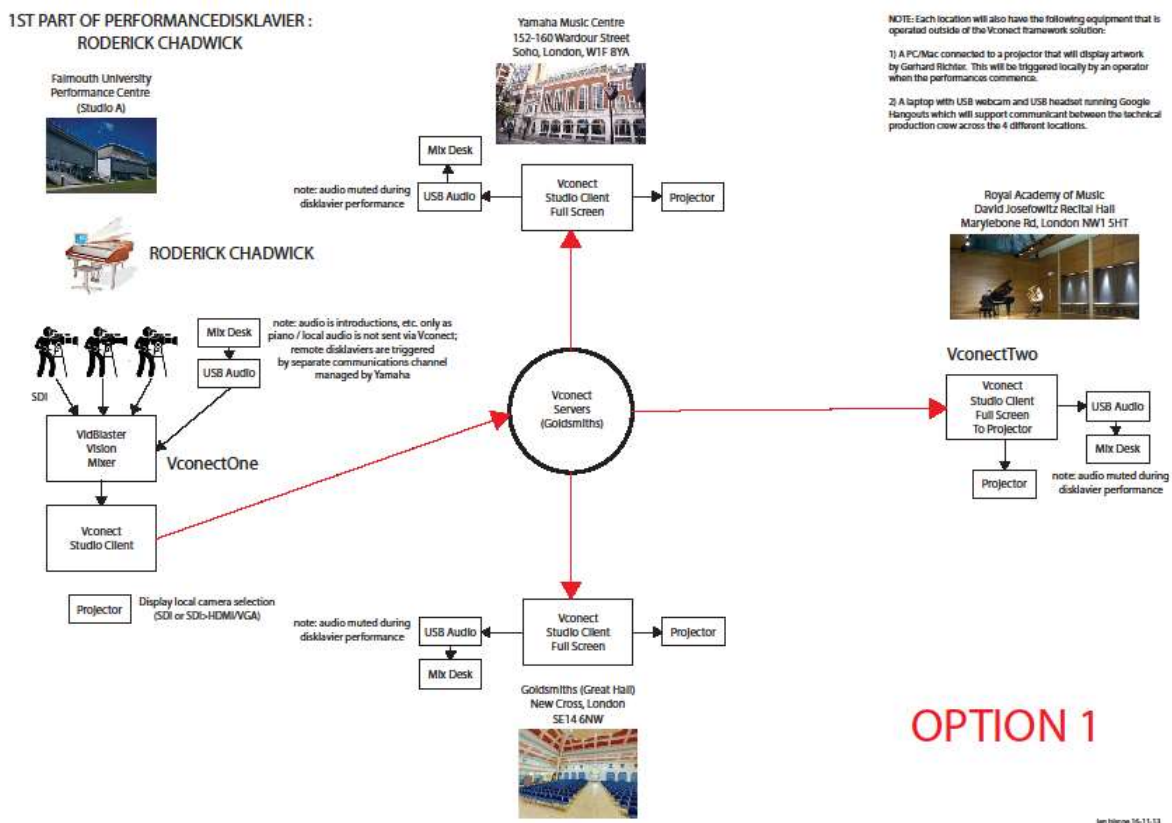
The various forms of Disklavier are essentially modern pianos that use electromechanical solenoids and optical sensors connected to LEDs allowing them to play notes and use the pedals independently of any human operator. In other words it is a modern version of the Pianola.

Sponsored by Yamaha, Falmouth University earlier this year staged a performance by Elton John where the artist played a Disklavier Grand Piano in Los Angeles, which resulted in Falmouth University's Disklavier being played remotely (including key's and pedal's mechanical activity) at its Performance Centre.

Associate Lecturer in Music, Jim Atchison, has composed a unique simultaneous concert to be performed on Yamaha's Disklavier pianos. The concert, inspired by the art of one of the world's greatest living visual artists, Gerhard Richter, will take place at Falmouth's Academy of Music and Theatre Arts (AMATA) and three other simultaneous locations using advanced Yamaha technology.

Taking place on Saturday 22 February 2014 from 7.30pm the simultaneous performance will occur between Cornwall and London, at Falmouth University's Academy of Music and Theatre Arts, the Royal Academy of Music, Goldsmiths University and Yamaha Music London.

The concert will involve four Yamaha Disklavier pianos separated by 300 miles and remotely controlled by a parent Disklavier at Falmouth's Academy of Music and Theatre Arts (AMATA), which will be played by just pianist Roderick Chadwick. In addition, the same music re-composed by Jim Aitchison for the strings of the Kreutzer Quartet, will be performed back from The Royal Academy of Music (RAM) to all the other venues via audio link.



The Video connection between audiences in four remote locations for this large-scale project is supported by Vconnect's Studio Vclients. Furthermore, using questionnaires, the audience response in the four locations will be explored.

April: Connected CAVE's audience evaluation

During a four week period in March and April Vconnect's CAVE's will be installed in two performance studios at the Falmouth University Performance Centre.

This installation is aimed to serve a several purposes:

- Experimentation by dance and theatre students with the remotely connected studios.
- Experiments comparing the audience response to a live performance (in one studio) with a remote audience watching the streamed (and interactive) version in the other studio. As a follow up to CWI's and Falmouth University's successful collaboration CWI's new GSR measurement system (capable of monitoring 40 participants simultaneously) will be deployed.
- Experimentation by Miracle Theatre as preparation for their performance of *The Tempest* (Shakespeare) in two remote locations (see below).

2nd Week September – The Tempest performed by Miracle Theatre in two remote locations using Vconnect's Studio Vclient.

Shakespeare's play *The Tempest* is set in two locations on a (mythical) island, one location is where the mortals live and the other where immortals dwell. Only occasionally will an actor cross the divide.

Sponsored by BT's Super-Fast-Broad-Band (SFBB) using Vconnect's Studio Vclient, there will be a performance using two connected theatres. In each theatre we will use large screens to stage a live connected performance.

This is a real "piece de resistance" for Vconnect's performance use case as well as for BT's roll-out of SFBB. Audience evaluation, possibly using GSR sensors, will be part of the event, of course.

9.2 Socialisation Use Case

Similarly to the remote performance of *The Tempest* being the large field trial for the Performance Use Case, a field trial is in preparation at SAPO-Campus aimed to take place in the June-July timeframe. Here the user response to Vconnect's Home Vclient will be evaluated.