Joint evaluation of communication quality and user experience in an audio-visual virtual reality meeting

Anders K. Møller[1], *Pablo F. Hoffmann*[1], *Marcello Carrozzino*[2], *Claudia Faita*[2], *Giovanni Avveduto*[2], *Franco Tecchia*[2], *Flemming Christensen*[1], *Dorte Hammershøi*[1]

[1]Section of Acoustics, Department of Electronic Systems, Aalborg University,
Fredrik Bajers Vej 7, 9220 Aalborg Ø, Denmark
[2] Laboratorio di Robotica Percettiva, TeCIP - Scuola Superiore S.Anna,
Via Alamanni 13B, Ghezzano, Italy

akm@es.aau.dk, pfh@es.aau.dk, m.carrozzino@sssup.it, c.faita@sssup.it
giova84@gmail.com, f.tecchia@sssup.it, fc@es.aau.dk, dh@es.aau.dk

Abstract

This paper presents a method for evaluating a communication situation in a multimodal multi-user virtual environment. In the evaluation both the quality of the communication and the user experience will be addressed.

Twenty-four subjects participated in the experiment which was carried out in a virtual reality CAVE. Two different sound presentations were used (Diotic and Binaural) and two different video presentations using two levels of video quality. The mix between these two variables created four different sessions. In each session the subjects observed a virtual meeting. The subjects tried all sessions and rated their experience on different scales. The subjects also answered a questionnaire about the content of the meeting.

The preliminary results from the experiment showed no differences between the subjects' rating of the experience. Tendencies of relation between visual cues and subjects ability to detect the emotional state of the meeting members were shown.

Index Terms: speech intelligibility, virtual reality, body language, telecommunication.

1. Introduction

In the last decade different virtual reality systems have been created where the users can communicate with each other in virtual worlds. Virtual reality systems are often multimodal systems where the users can, for instance, both see and hear each other. Speech perception is multimodal as well and can be influenced by visual information. One example is the McGurk effect [1] that shows how lips movement can influence what we hear.

In a communication situation it is not only important to understand the words being said it is also important to understand the meaning. One sentence can be understood in two different ways depending on whether the speaker is joking or being serious. It is important for a listener to be able to detect whether a person is, for instance: happy, angry, joking or serious. These emotional states are often expressed through our voice, our facial expression, or our body language - also called the indexical characteristics [2].

Talking agents and avatars have proven the use of visual information to improve several aspects of a conversation. Talking agents which add body language, facial expression and lip-synchronized articulation have proven to increase intelligibility of synthetic speech [3]. Other virtual reality studies have shown that visual aspects like more realistic looking avatars can affect the perceived quality of communication in virtual reality [4]. It is very likely that the introduction of more detailed recordings of real humans might also improve the perceived quality of the communication.

In the framework of the EU BEAMING research Integrated Project previous experiments have evaluated the suitability of remotely shared virtual environments (SVE) or mixed reality environments (MRE) for communication purposes. BEAMING-based systems have been implemented in order to allow actors and directors to meet in an SVE [5] or a MRE [6] in order to rehearse scenes for a play or a movie, that is, to perform dialogues and blocking (positions, movements, and displacements of actors in the scene) rehearsal.

Even though methods exist for testing communication between users in an audio-visual contexts [7] they often focus on recognizing single words in fixed sentences. When talking, the context of the conversation

10.21437/PQS.2013-27

facilitates word recognition in a sentence [8]. Instead of focusing on recognizing single words from fixed sentences this work will introduce a more holistic approach to measuring the quality of a communication in a multimodal virtual environment. In this paper we will focus on a case of a mediated long-distance meeting.

The aim of this study is to evaluate whether spatial sound can help the subjects to keep track of who is saying what during a meeting with multiple members. The use of spatial sound have previously proven to help users to locate visual objects and also improve users sense of presence in audio-visual virtual reality [9]. The present study will also investigate whether the visual rendering of the meeting members makes it easier to follow the conversation and understand the behavioural characteristics of the members.

The evaluation is carried out with a focus on evaluating the conversation including the rating of the indexical characteristics. The relation between the quality of the audio-visual rendering and the sense of actually being there will also be addressed.

2. Methods

The method for evaluating the communication situation will be based on an experiment using recordings of a meeting to present a remote meeting. In the experiment subjects will act as a secretary participating in the meeting. Even though no real interaction between the actors and the subject exists the recordings will include situations where the subject is addressed by the actors during the conversation.

2.1. Session description

Four sessions were recorded for the experiment. Each session was represented as a meeting between two actors. The actors were not professional actors and had no notable experience with acting. The secretary is observing the remote meeting between the actors. It is the secretary's job to follow the conversation between the two actors and take notes.

The topic of the meeting for each session was the same but it changed in small particulars, while the emotional attitudes of the actors alternated from one session to another. The elements, variables and constants of the session are shown in table 1. The constants are the things keept constant across the sessions and the variables are the things that changed. In order to create the same level of interest and attention in each session the general topic, the time of each session, and the expression of emotions (which alternated between the actors in different sessions) were the same. In the first part the actors' quarrel overlapped each other's voices, in the second part the actors reconciled and laughed together, in the third part they address the subject speaking about him and pointing at him with their fingers. In the final part of the session, after a brief digression, the actors interrupted the conversation for a few seconds then they made an agreement about the next meeting and greeted. Each session took around seven minutes.

The actors took on two different behavioral characteristics, the aggressive and the thoughtful behavior, which they switched between the sessions. In order to have a similar expression of emotion two males were chosen as actors. People of the same sex do not just have a more uniform voice they also have a more comparative level of the body communication [10]. Two actors of the same sex was therefore used.

Constants
Overall Topic: Virtual Environment Experiment **Sessions Duration:**around Seven Minutes each **Emotion expression Four parts:** 1-contrast/overlap, 2-agreement/laugh, 3-interaction, 4-suspension
Variables
Subtopic: Social Phobia in, Flight training (with HMD or CAVE technology) **Actor Character:** different in every session

Table 1: *Overview of constants and variables for the four sessions*

2.2. Setup and equipment

A virtual meeting room has been created with a table surrounded with chairs using the 3D computer graphics software 3dsMax.

Four video sessions were recorded of the two actors.

The visual part of the video was realized through the live-recording of the stream coming from a Microsoft Kinect, which was real-time converted to a polygonal mesh and subsequently compressed and stored in a binary data file. This was achieved by means of a software module allowing also to manage playback and to vary different parameters influencing the quality of the visual rendering, namely the bitrate and the quantization levels. The module has been implemented as a plug-in for XVR [11], a VR integrated development environment allowing to author and manage complex virtual environments in a wide range of installations.

The experiment was taking place in X-Cave. X-Cave is an immersive visualization system developed by PERCRO Laboratory of Scuola Superiore Sant'Anna and located in its facilities in Pisa. X-Cave is a large-sized CAVE-like system (4x4x3 m) composed of 4 walls, each divided into several tiles in order to ensure a good resolution. Front, right and left screens are subdivided into four back-projected tiles each, for a global resolution of about 2500x1500 pixels per screen, whilst the floor is subdivided into six front-projected tiles for a global resolution of about 2500x2250 pixels. Globally the system includes 18 projectors and an Intersense IS-900 head tracker. The X-CAVE is managed by means of the XVR technology, exploiting its distributed rendering features [12] on a cluster of five workstations. In the CAVE the meeting room and the actors are presented with a stereoscopic image on the walls and the floor. The sound is played through an audio software developed during the EU BEAMING project with the purpose of transmitting 3D sound over networks [13]. The sound was presented to the subjects through equalized Beyerdynamic DT 990 Pro headphones using a Roland QUAD-CAPTURE sound card.

The binaural processing which is used in the experiment include usage of digital filters known as Head Related Transfer Functions (HRTFs). The HRTF database in this experiment is based on 2 degree angular resolution measurements of the ”Valdemar” dummy head developed at Aalborg University [14]. The two actors are rendered in the respective starting positions of the actors, which didn’t change during the experiment. The position and head-rotation of the subjects are tracked and used for the binaural sound rendering.

In the experiment the subject is placed on a chair on one side of the virtual desk and the two actors are positioned on the other side of the desk. One actor is located approximately 45 degrees to the left of the subject and the other approximately 45 degrees to the right of the subject.

The subject is wearing 3D glasses, a head tracker and headphones during the experiment. Fig. 1 shows a picture of a subject in the setup wearing all the equipment.

2.3. Design

The evaluation method is based on a within subject design with two independent variables. The first variable is the audio presentation which varies between a mono processing method where the participants receive a diotic sound presentation (presenting the same stimulus to both ears) and a binaural processing method [15] creating a 3D sound experience. The other variable is the visual quality which varies between one having the best visual quality (high bitrate and low quantization levels) and one having a worse quality (low bitrate and maximum quantization levels). A session was created for each combination of the two variables giving a total of four different sessions:

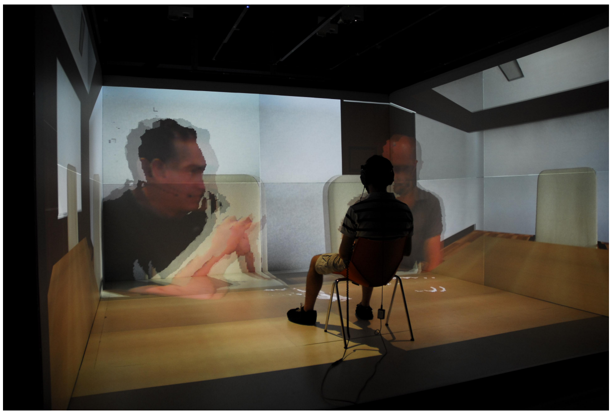

Figure 1: *A picture of the setup in the CAVE. The image looks blurry due to the stereoscopic image*

- Session A: High quality video, Binaural sound
- Session B: High quality video, Diotic sound
- Session C: Low quality video, Diotic sound
- Session D: Low quality video, Binaural sound

The order in which the sessions were presented to the subjects were balanced.

2.4. Measurement

After each session the subjects were given a questionnaire. The questionnaire contained 10 questions (Q1-Q10) related to the content of the meeting and stated as multiple choice questions with two options, beside one question which had three options(Q7). The questions would either address the content of the conversation or the actors with questions like: ”who prefers the modelling of a more playful environment?” or ”when will the next meeting be held?” Two questions were also related to the behavioral characteristics of the actors.

As already mentioned the subjects were told to take notes during the meeting and the subjects were allowed to use the notes when answering the questionnaire. The use of notes served different purposes. First of all it was a natural part of the secretary role. It also introduced extra head movement which could affect the sound experience because the subject would most likely turn the head to take notes. Finally it helped the subject when answering the questionnaire. By allowing the subject to use the notes we tried to change the task from a memory task to a task of following the conversation. If we had given the subjects the questions to answer during the meeting we would risk that the subjects would only focus on listening for the specific questions and not experience it as a real meeting. We could not base our measurements on the notes alone either because people use different techniques when taking notes, meaning some will write a lot and others less which would greatly influence the results.

The questionnaire also contained five questions asking the subjects to rate how they experienced each session. The five questions were stated on Visual Analog Scales (VAS) with bipolar end-labels. The scales were later translated to a linear scale from 0-15.

- Q11: During the experience, how often did you think of yourself as actually participating in the meeting? (Not very often -Very often)
- Q12: How difficult was it to follow the conversation during this part of the meeting?(Very difficult-Not very difficult)
- Q13: To what degree did you feel tired during this part of the meeting?(Not very tired - Very tired)
- Q14: How do you rate the quality of the video in this part of the meeting?(Not very good - Very good)
- Q15: How do you rate the quality of the audio in this part of the meeting?(Not very good - Very good)

The reason for asking the subjects whether they felt tired as in question 13 was that if they felt tired it might have been because they had problems following the conversation and therefore had to focus more, or they felt tired because they did not feel that they were really involved in the meeting.

After completion of all session the subjects were also given a questionnaire with two open questions to each session where the subjects could comment on the audio and video rendering and how that affected their experience.

The purpose of the open questions were to get qualitative data which could provide further insights into potential findings.

2.5. Procedure

At first the subject received instructions about the experiment. After the instructions the subjects were told to place themselves in a chair positioned in the middle of the room. They received help with putting on the equipment. The subjects were also provided a block of paper for the notes. After checking the equipment was placed where it should the first session started.

After each session the light was turned on and the subjects were told to answer the questionnaire. If the subject needed a small break after the session this was allowed. The subjects completed four sessions each, followed by a questionnaire. After the last session the subjects were given an additional questionnaire with open questions related to their experience of each session.

The whole experiment were carried out in Italian.

2.6. Subjects

Twenty-four people volunteered to participate in the experiment 18 males and 6 females. The subjects were recruited both externally and internally with the majority recruited internally. The youngest subject was 25 and the oldest 41 with a mean age of 29.2. The subjects all had Italian as their first language.

3. Preliminary results

The number of correct answers in the questionnaire for question Q1-Q10 for each subject for each session were counted. Unanswered questions counted as errors. In table 2 the number of correct answers for each session are presented. In session A and B the total number of correct answers are a bit higher than in session C and D. Session A and B are the two sessions with the high video quality where C and D have the low video quality. Fig. 2 shows how many of the 24 subjects that answered each question correct. The results from each session is presented in separate columns. In the answers to question 8 and 9 there are some interesting variations between the sessions.

Session	A	B	C	D
Score	220	226	198	199

Table 2: *The score corresponds to the total score of correct answers (out of 240) in Q1-Q10 questions for the 24 for subjects.*

- Question 8: Who seems to have the most aggressive behaviour?
- Question 9: Who seems to be the most thoughtful?

If the answers to question 8 and 9 in session A and B is compared with the answers in C and D we see a difference in the number of correct answers. Both question 8 and 9 address the attitude or the behaviour of the two meeting members.

In some questions the number of correct answers are close to 24 for all four sessions meaning that some of the questions might have been to easy to answer.

The subjects were told to rate each session based on 5 different scales. In table 3 the mean for each rating

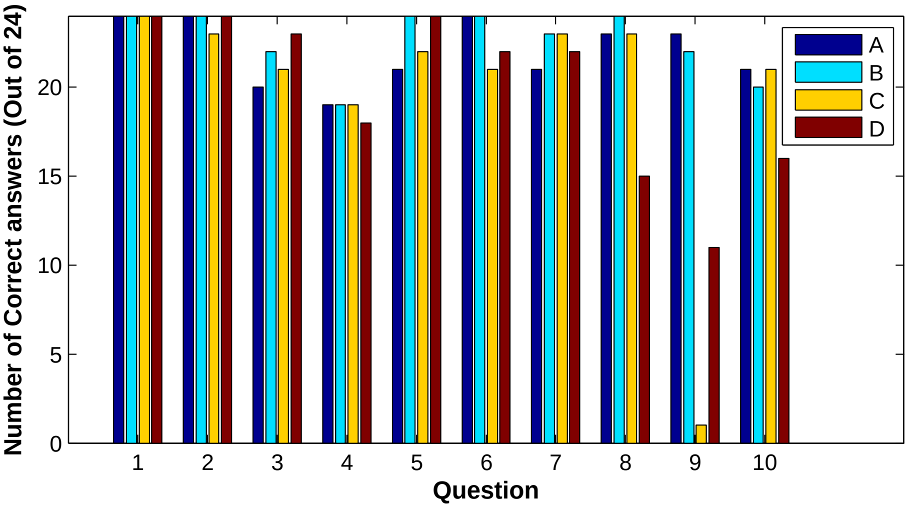

Figure 2: *Overview of the number of subjects that answered correct for each question in Session A, B, C, D*

Q	A		B		C		D	
	M	SD	M	SD	M	SD	M	SD
11	3.28	2.34	3.17	2.22	4.27	3.41	3.62	2.84
12	9.12	3.74	9.99	3.80	11.44	2.94	10.80	2.72
13	5.23	4.01	4.64	3.04	4.37	3.77	4.33	3.03
14	4.68	3.53	4.50	2,98	3.25	2.26	3.29	2.31
15	5.77	4.09	7.45	3.18	8.96	3.97	8.38	3.16

Table 3: *The mean and standard deviation of the subjects' rating in question 11-15 ratings shown for all four sessions: A(HQ video + binaural sound), B(HQ video + diotic sound), C(LQ video + diotic sound), D(LQ video + binaural sound)*

is presented along with the standard deviation. While Question Q14 and Q15 is a measure of the perceived quality of the independent variables (video and audio) Q11-Q13 is the rating of the experience of the sessions which is varied by the two independent variables. Question Q11 addresses whether the subject felt they were actually participating in the meeting. Question Q12 is a measure of the subject's ability to follow the conversation. Question Q13 is a rating of whether the subjects felt tired after experiencing each session. Question Q14 was a rating of the video quality and it was therefore expected that session A and B would be rated higher than session C and D since A and B was actually recorded in a better quality. Question Q15 was the rating of the audio quality where spatial sound were provided in session A and D. It was therefore expected that these two sessions would be rated higher than B and C.

After the subjects had completed all sessions they were given two open questions to each session where they could address in what ways audio and video affected their experience.

In the open questions several subjects pointed out an asynchrony between audio and video. Several subjects also reported differences in the video quality and audio quality. The comments about the video quality were mainly related to the overall image quality a few subjects also pointed out issues with the recording of the actors like: ”a small part of the actors body disappeared” or ”the actors appeared to big”.

In relation to the audio it was mentioned that the actors sometimes ate their words, and that if the audio had been louder it would have been easier to follow the conversation.

4. Discussion

As mentioned in the results section, several subjects commented on the asynchrony between audio and video which could have masked the intended effect of the difference in the audio and video presentation if this was the case it would not just affect the ratings of audio and video quality in question Q14 and Q15 but also affect the experience ratings in Q11-Q13. The problem with the synchronization occurred because the video and the audio were presented using two different software programs and sometimes a missing data package could delay the video stream creating asynchrony between video and audio. Even though this asynchrony did not become prominent before in the end of the sessions it could still have effected their judgement.

Despite the mentioning of poorer or better audio quality in some of the sessions nobody mentioned the spatial properties of the sound. This could be because they never really experienced the spatial properties or it could be because they were never explicitly asked about the spatial properties. If the subjects had been asked directly it would have compromised the purpose of the experiment making them more aware of the spatial properties of sound.

It was shown that in the question Q1-Q10 there were few errors when answering the questionnaire which could indicate that it was not sufficiently difficult to answer the questions, making the spatial properties indifferent for the task. Even without the spatial properties of the sound it could have been possible for the subjects to distinguish between the two actors based on their voices. It might be necessary to create a more challenging task for the subjects by either having more overlapping speech or extra actors to actually show an effect of spatial sound.

The choice of having only two actors was forced by technical limitations that limited the frame rate. If we had used more actors it would have introduced even more latency. At present a solution is being developed which will allow the management of more than one Kinect stream without performance penalty in order to have more actors. This could for instance be four actors instead using two Kinects. This will also enlarge the useful spatial area around the subject and increase the angles from which the actors could be positioned according to the subject. The angle from which the sound sources could come would thereby also be increased. This, together with the improvement of the synchronization, would probably present a more challenging task better suited for evaluating spatial audio.

In this paper a method for evaluating the quality of a communication situation in a multimodal virtual environment has been suggested. In order to draw any final conclusions about the effect of spatial sound and video quality on a communication situation, the problem with the synchronization will have to be fixed and it might also be necessary to add extra actors to increase the difficulty of the task.

5. Acknowledgement

The work presented in this article is supported by the EU FP7 BEAMING project, project number: 248620.

6. References

[1] Macdonald, J., and McGurk, H., “Visual influences on speech perception processes”, Perception and Psychophysics., 24(3):253–257, 1978.

[2] Goldstein B. E., “Sensation and Perception.”, Seventh edition: 311–327, 2007

[3] Beskow J., and Elenius K., and McGlashan S.,“OLGA - A dialogue system with an animated talking agent”, Proceedings of Eurospeech,1997

[4] Garau M., and Slater M., and inayagamoorthy V., Brogni A., and Steed A., and Sasse M. “The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual”, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Pages 529-536, 2003 Environment

[5] Normand, J., and Spanlang, B., and Tecchia, F., and Carrozzino, M., and Swapp, D., and Slater, M. “Full body acting rehearsal in a networked virtual environment - a case study”.Presence: Teleoperators and Virtual Environments 21(2), 229-243 (2012)

[6] : Steptoe W., and Normand JM., and Oyekoya O., and Pece F., Giannopoulos E., and Franco Tecchia, and Steed A., and Weyrich T., and Kautz J., and Slater M., “Acting Rehearsal in Collaborative Multimodal Mixed Reality Environments”. Presence: Teleoperators and Virtual Environments 21(4), 406-422 (2012)

[7] Cooke, M., and Barker, J., and Cunningham, S., and Shao, X. “An audio-visual corpus for speech perception and automatic speech recognition”, Journal of the Acoustical Society of America., 120(5): 2421–2424, 2006.

[8] Meyer, D. E., and Schvaneveldt, R. W.,“Evidence of a dependence between retrieval operations.”Journal of Experimental Psychology.,90(2): 227–234, 1971.

[9] Larsson P., and Vastfjall D., and Kleiner M.,“Better presence and performance in virtual environments by improved binaural sound rendering”, AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, 2002)

[10] Argyle M., “Bodily Communication”, Methuen, London, 1988

[11] Tecchia F., and Carrozzino M., and Bacinelli S., and Rossi F., and Vercelli D., and Marino G., and Gasparello P., and Bergamasco M., “A Flexible Framework for Wide-Spectrum VR Development.” Presence-Teleoperators and virtual environments 11(19), 2000

[12] Marino G., and Vercelli D., Tecchia F., Gasparello P., and Bergamasco M., “Description and Performance Analysis of a Distributed Rendering Architecture for Virtual Environments”, 17th International Conference on Artificial Reality and Telexistence, pp.234-241, 2007

[13] Madsen E., and Olesen S. K., and Markovic M., and Hoffmann P., and Hammershi D., “Setup for Demonstrating Interactive Binaural Synthesis for Telepresence Applications” Forum Acusticum, 2011

[14] Bovbjerg B. P., and Christensen F., and Minnaar P. and Chen X. “ Measuring the head-related transfer functions of an artificial head with a high directional resolution”,Proceedings of 109th Audio Engineering Society Convention, 2000

[15] Møller, H., “Fundamentals of binaural technology.” Applied acoustics., 36(3): 171–218, 1992.