NTNU – Trondheim
Norwegian University of
Science and Technology

Telepresence Quality

Daniel Puig Conca

Master of Telematics - Communication Networks and Networked
Submission date: June 2012
Supervisor: Leif Arne Rønningen, ITEM
Co-supervisor: Otto Wittner, UNINETT

Norwegian University of Science and Technology
Department of Telematics

Problem Description

Name of student: Daniel Puig Conca

One of the goals of a telepresence system is to make a user feels as he/she was present in other remote locations. The aim of this thesis is to select test methodologies and carry out tests of the user's perception of the telepresence experience.
There are different parameters that are closely related with the perceived quality, one of the most important is the delay, which will be dealt deeply.
A comparison between face-to-face stimulus-response tests and system-to-face will take place for concluding the degree of near natural perception is experienced by a user.
Thus, a test paradigm will be performed for testing perceived quality of video and sound collaborations.
Finally, a real case scenario with musicians and a conductor playing together will be done.

Assignment given: January, 16th 2012

Supervisor: Leif Arne Rønningen, ITEM

Abstract

Nowadays, one of the aims of telepresence systems is to provide a sensation of nearness to people who are interacting with this type of systems. Many factors have a relevant repercussion in providing this feeling and some aspects are more important than others, depending on the scope of use. This thesis presents several studies made in order to analyse the degree of importance each factor has.

One of these factors treated is the delay which limits the interactivity. For this reason, in this thesis, a method is proposed to measure the delay through a telepresence system. Another factor treated has been the frame rate in order to figure out which is its influence. In addition, an stereoscopic 3D setup was performed to analyse the degree of perceived depth which was introduced into the system.

Finally, several pilot tests focused on musical rehearsals were made to evaluate the influence of the delay. The recording was made at 60fps in a high-definition quality. Subjective opinions about the interactivity and perception of this sort of systems were gathered.

It was concluded that this sort of system was viable for interactive applications like conducting a choir, but an effort must be done when decreasing the amount of delay added by end devices. In fact, the conductor tolerated a delay (round-trip) about 118ms in rhythmic music, being still possible to conduct with difficulties. In contrast, the delay tolerance increased up to 160ms when conducting a more melodic piece of music.

However, the use of 3D when there is more than one viewer does not produce much benefits. Instead of that, it is proposed to analyse multi-view systems as a future research.

Preface

The decision of doing this project was taken in great measure for the advices and the new acquired expectations during a course lectured by Leif Arne Rønningen.

First of all, it must be said that this project has involved a large number of people. The help given by the ITEM institute at NTNU was decisive. All the equipment needed was provided by these institutions as well as a lab where this research was developed.

We must be thankful to Pål S. Sæther who has been very patient and helped us to get all the material needed. Besides this, he also contributed to build the test environment that was necessary to perform the tests.

During the response time tests, Andrés Cervantes, Diego Salvador and Elaheh Vahidian participated, thus, we really appreciate their efforts. Special thanks to all of them.

We would also like to express our gratitude to Otto J Wittner for his cooperation and the equipment provided from Uninett which has been used in this work.

We will always be grateful to the group of members of *Nidaros domkor* (Anita Brevik, Anne Sigrid Imsen, Solveig Meland, Mariel Eikeset Koren, Terje Aandalen, Jon Bang, Martin Eikeset Koren, Lars Sydnes). All of them were necessary and indispensable and made possible this research. They were the engine of this study as well as Torleif Weydahl. Vivianne Johnsen Sydnes was one of the most relevant participants. She was the conductor of the choir and she was very kind of taking part in this project.

In addition, a big thanks to Miriam Navarro Conca for looking through this thesis. Her suggestions improved the writing of this research.

Finally, we really appreciate the work of Leif Arne Rønningen for being such a persevering teacher with his dedication and attention during this stage. He provided us advice and practical help. He has been an unconditional support whenever we needed him.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Current telepresence systems are used by many companies as a way to reduce costs. This option involves not only reducing expenses but an increase of the participants comfort as well as a reduction in travelling time. These and other advantages are provided by this sort of systems. Furthermore, the possibility of using them in other activities is being explored.

Nonetheless, all activities do not have the same requirements in terms of technical specifications. Thus, each one must be analysed to overcome their claims.

The aim consists in finding the parameters' tolerance that provides an optimal user experience. With this information, the technical goals are fixed in order to carry out a technical implementation.

The Thesis

This work presents several parameters that determine some aspects of the perception, interactivity, feelings and so on, by using these systems. Due to the fact that this topic is very broad, the content of the thesis is focused on some of them.

Consequently, in order to study the delay, a simple method to perform measurements is proposed as well as new ways to improve their performances as future researches. This method is the tool used to study other aspects as the frame rate.

A study of the frame rate and its relation to the response of a stimulus was done. The collaboration of people was necessary to measure their reactions.

As a practical case, a way to implement real-time 3D in these type of systems is described. Furthermore, the consequences that produces this fact are explained.

Finally, a telepresence system was built up to carry out a musical rehearsal. In order to evaluate its performances, this platform was used for the collaboration between members of a choir and a conductor. Several tests were performed to try to test the user's perception of the telepresence experience as well as to analyse the viability of this type of initiatives. A co-supervisor and a choir conductor cooperated in this project, signing a collaboration contract before the start.

Scope

The study made about the relation between the frame rate and the response time must be taken as an starting point for a wider research. The limitations are imposed by the number of people who participated in the tests. In this case, there was about a hundred samples taken for each scenario from four collaborators. Consequently, the conclusions cannot be generalized.

Respecting the subjective tests performed over the system built, some considerations have to be kept in mind. The data collected are collaborators' personal impressions. In addition, the number of people that tried the system was limited. Therefore, there is a big disparity in some points. Besides this, the threshold obtained refers to the conductor's feelings. Consequently, not everybody could tolerate the same delay.

Work

Most of the emphasis on this thesis has been centred on practical work. One of the goals was to get a running system in which to perform several tests. Its work was evaluated as well as its effectiveness.

The first steps were taken from the equipment available in the laboratory. Furthermore, the next work was determined by their capabilities. Besides this, the subsequent acquired equipment was subordinated to this choice. In fact, working with the new equipment implied to make a thoroughly understanding of how it worked to be able to judge the results and also to set the proper configurations.

Apart from this work, some aspects of this thesis are focused on matters belonging to other disciplines. The topic presented in Chapter 4 has more relation with psychology.

Due to the limitations on time imposed by the delivery time of the equipment, some of the topics have been developed in line with the study. Instead of this, other topics have been waiting until they had the opportunity to research on them because they needed the collaboration of other individuals.

However, the structure of this thesis is presented sequentially, as it is shown in the paragraphs below.

Thesis organization

The work presented in this thesis has been divided into several topics which make up the different sections. Chapter 2 presents some technical parameters related to the perceived quality of a telepresence system. Furthermore, researches about the use of this sort of systems in musical performances are presented. In Chapter 3 the methodology that has been used to obtain data is explained and analysed. Two parts are clearly distinguished. The first part deals with the delay. An overview of several researches that have taken place in recent years concentrated on the study of the delay in telepresence systems is exposed. Continuing the study of the delay,

Section 3.2 presents a method for measuring the optical delay in a telepresence system, that is to say, since a fraction of light is captured by a camera on one end and it is projected at the destination. Additionally, the principles of this method are also used in the following chapters. In the second part, Section 3.3, general methods to perform qualitative analysis are outlined. One of these methods was used to gather subjective impressions of the participants who tested the developed telepresence system. Chapter 4 deals with a characteristic parameter of the video recording, the frame rate. The relation between this parameter and the reaction time of an individual is presented. The goal was to find out which degree of correlation exists between them. In Chapter 5, a brief description of current active stereoscopic 3D is shown. In addition, a way to implement this feature in real time is presented, using the equipment used in the previous and future sections. The idea was to evaluate the performance and the sense of depth experienced. The developed system and the testing sessions are explained in Chapter 6. Several pilot tests were carried out in different scenarios, obtaining the views and ratings of participants. The next two chapters constitute the conclusion of the work. A discussion about the results and statements described is done in Chapter 7. The last chapter presents a brief general conclusion.

Supplementary, there are seven appendices attached. Appendix A shows an outline of equipment used. The graphics regarding the Response Time results are presented in Appendix B. Appendix C clarifies the depth cues. A proposal implementation to delay video is proposed in Appendix D. The scripts used to delay video are presented in Appendix E. Furhtermore, the results of musical rehearsal interviews are shown in Appendix F. Finally, in Appendix G the permission contracts are attached.

Chapter 2

Telepresence Quality

The goal of a telepresence system is to make feel users as they were at the same place. Therefore, the quality of the system has to be near-natural. As it is known, the quality is dictated by the demands of the user. The users set which degree of tolerance is acceptable.

However, in order to decide what values are tolerable, the equipment and network-traffic parameters allow the tolerance.

2.1 Equipment parameters

2.1.1 Video

Quality of a video is imposed by the resolution and frame rate. The more resolution, the better the detail will be visible in an image whereas frame rate is the number of frames that are shown in a second. Consequently, fast movements will be clearer, avoiding the 'ghost effect'.

2.1.2 Sound

Sound quality is characterized by bit rate, and compression used. Moreover, the inclusion of more audio channels will generate a more reliable perception, providing that the configuration and placement are correct.

Moreover, the synchronization between these two channels of information is essential. If the participant can notice an incoordination, the feeling of nearness is broken.

2.1.3 Delay

The process time that an equipment needs to do its functionality is an important factor that contributes to the global delay of the entire system. In subsequent chapters, the importance of the delay introduced by the equipment is analysed.

2.2 Network parameters

The quality of a telepresence system is also thoroughly related to the parameters presented below. Each one affects a different perception aspect, being some of them related.

2.2.1 Packet loss

Telepresence systems provide a service which has real-time priorities. This fact forces to use UDP (or similar protocols) as a transport protocol. Consequently, the loss of packets can occur and these packets will not be retransmitted because they have to arrive within the right time slot. That is to say, the real-time streams must stay on time.

Hence, packet loss diminishes the perceived quality. Nonetheless, this loss can have a big or a small impact into the systems depending on its implementation. For instance, this effect will be aggravated if incoming packets are based on previous packets. The loss will produce a chain reaction. The Intra-frame coding for video is performed just regarding the information on the current frame. However, there are other codings that use information of previous frames in order to construct the new frames. As an advantage, this last method is bandwidth-saving. In fact, the choice of a video codec is an important design decision.

Because of this loss, strange effects can appear like blur, image freezing and so on.

2.2.2 Delay

The delay is the time that a packet spends between one end of the system to the other end. The quantity of delay that introduces the entire system is quite critical in some applications. The interactivity and nearness feeling of the system will be set by the delay. In posterior sections, this parameter is analysed deeply.

2.2.3 Jitter

Jitter or variation in delay is another important parameter that appears because of network congestion. This effect can provoke that a packet misses its playing-time slot. In order to counter this fact, buffers are used. Nevertheless, the introduction of this mechanism contributes to the latency. For instance, Cisco recommends to use a jitter buffer of less than 10ms.

2.2.4 Bandwith

It is not a direct-connected parameter to the perceived quality. Bandwidth limits the among of data that can be transmitted. As a consequence, the quality of the video and the number of sound channels are subordinated to this parameter.

2.3 Current telepresence systems

Current telepresence systems are mainly focused on providing a tool for oral communication at a distance. In fact, they present communication delays about 150 and 250 ms in the best cases. This fact occurs because they are paying particular attention to this area in which these values are acceptable.

Nonetheless, new researchers are focused on extending the possibilities of these systems. Consequently, the quality standards are subordinated to the demanding requirements of these paradigms.

2.4 Musical performances

This is one of the main focused point of this thesis so that a group of researches that treat this topic are presented.

One the most important projects that have been done is the Distributed Video Production (DVP) (Konstantas et al., 1997). This project consisted in the performance of a distributed musical rehearsal (Konstantas et al., 1999) through an Asynchronous Transfer Mode (ATM) network. A subjective and an objective evaluation was done. The results can be seen in (Orlarey et al., 1998).

Another important group of researchers is grouped within the Distributed Immersive Performance (DIP) project. This work started in 2002, being continued later on. In (Chew et al., 2004) is presented an overview of several experiments that were performed. The idea consists in creating an immersive technology for distributed musical collaboration.

2.4.1 Other paradigms

Internet has been shown in this text as a transmission media by which signals are transmitted. Trying this environment does not affect the different systems that use this network so far. As a curiosity, other ways of thinking have emerged trying to live with the features of this medium. Golo Föllmer says that what you have to do is to find a paradigm that works with the features of this media or, in other words, find a paradigm that fits the media.

Within this idea and taking advantage of the delay provided by the network, an experiment (Handberg et al., 2005) between the Royal Institute of Technology in Stockholm (KTH) and the Stanford University in Palo Alto was carried out. The project consisted in conducting several concerts of improvised music between two nodes located at each university. At each node two musicians were playing as well as an audience was present. In this case, a delay about 250ms was achieved. The key experiment was not trying to play music with the best timing possible, but the musicians adapted themselves to the network. This network was used as a method of introducing a sound effect. Note that the concert not only was transmitted but also the stage where the audience was attending the performance.

Chapter 3

Methodology

In order to perform measurements, two approximations can be taken regarding the type of data that is wanted to be gathered. Therefore, quantitative methods involve the analysis of numerical data. As a result, an objective conclusion is obtained according to the measurements. For instance, this data could have been obtained using a proper measurement method. In other case, data collected from surveys is analysed, conducting a statistical analysis over it. Qualitative methods deal with descriptions. The data obtained cannot be measured. Instead of that, statements and interpretations about what is observed are done. Subjective data is now stated.

Section 3.1 and 3.2 presents a quantitative method to measure the delay. After that, two methods to acquire qualitative data is exposed in Section 3.3.

3.1 Previous work on delay

The delay is one of the most important parameters that characterizes a telepresence system due to the fact that it will set the level of perception that the user has of the system. It defines the sense of distance that a user feels with respect to its namesake in another location, in other words, the degree of presence.

The delay, even when a person does not realize, is a daily parameter that has a relevant influence in our lives. When talking, the mechanical waves emitted by our vocal apparatus take about 3ms to travel one meter [1]. Nevertheless, oral conversations can be kept at different distances or, in other words, different delays can be tolerated. After that, depending on the area that is being focused, different delay values will be allowed, resulting in a degradation of the user's experience if these minimum values are not met.

[1] speed of sound in dry air at 20°C and 1 atmosphere pressure is approximately 343.2 m/s

3.1.1 The delay in oral communications

Oral communication is a process whereby the communication with other people is performed and it is composed of different stages.

The starting point refers to the conception of an idea that wants to be presented to people for keeping a debate or conversation. Afterwards, the physical process of talking takes place to transmit an acoustic waves-shaped modulated idea. At the same time that the information is transmitted and received, the other individual needs to stay alert to be able to receive the content and decode it. In order to continue with the conversation, the individual who has received the message must process it and think of an answer, repeating the whole process but this time starting from the other individual.

As it is well-known, conversation is a time-consuming process. An individual does not reply immediately after receiving the information. It is a slow process.

Therefore, the intrinsic features of an oral conversation cause tolerable latency values, which are higher when they are compared to other areas. Consequently, in (Bartlett, 2007) it is said that in the oral environment

> *"most humans don't notice audio delays of less than 150 ms, so this is the well-accepted one-way maximum latency in the voice environment"*

and in the opposite case

> *"most human notice delays above 250 ms"*

3.1.2 The delay in a musical performance

As it was mentioned above, the level of tolerance in the delay is determined by the scope of use. More restrictive degree of tolerance appears on the musical performance scope because of musical paradigm or interpretation requirements which are totally different from oral communication. It is needed an instantaneous response for a proper synchronization among musicians who are playing at the same time.

3.1.2.1 Physical delay on musical environments

In order to have an idea of delay values in this context, some examples are presented below.

> *...on a large stage at a symphony concert, diagonally located players on the edges can have a distance of 30 m, and about 100 ms for propagation of sound is needed*

(Gu et al., 2005, p. 87)

> *...the distance between a first violin player and a double bass player in a full-sized symphony orchestra (both sitting at the fifth row when counted from the conductor's point), is about 30 metres. A delay between the conductor and a trumpet player (sitting at the back row of the orchestra) is in the order of 46 milliseconds*

(Kleimola, 2006, pp. 3-4)

> *Each meter between source and microphone increases latency by approximately 3 ms.*

(Kleimola, 2006, p. 4)

3.1.2.2 Threshold on a musical performance

Numerous experiments have been carried out in order to determine the maximum amount of delay, which can be tolerated among musicians playing in sync.

In the research leaded by Schuett (2002) provides that the Ensemble Performance Threshold (EFA) to rhythmic music has a value between 20-30ms or, in other words, if musicians experience a delay between these values, the musical performance can be carried out without any problem.

A subsequent research supervised at the Stanford University by Chafe et al. (2004) has established new thresholds. In this research, an experiment which consisted of 2 visually isolated individuals clapping together in 2 different rooms was run with the intention of keeping up the rhythm. At the same time different delay values were injected to the received acoustic signals by each individual under test.

Three different cases were observed:
When the delay was less than what a person normally experiences, the tempo tended to accelerate. Once this exceeded a certain threshold, they obtained the opposite case, deceleration. Consequently, a range of values was found at which the interpretation was performed correctly.

When several musicians play a piece of music together in which different melodies are making counterpoint to finally compos a homogeneous melody. Timing is crucial and the correct length of the notes is also very important.

For example:

> *A 16th note at the tempo of 120 beats per minute (bpm) lasts 62.5 ms*

Whereby an out of time entry of more than 62.5ms produces a fault, and a corresponding musical desynchronization. When this is applied to a musical performance in systems that introduce delays, a threshold appears. This threshold must be taken into account.

The threshold is deeply related with tempo, for instance, in (Barbosa, 2002, pp. 112-114) an experiment is mentioned. This experiment took place in 2004 by the author's research group of that dissertation. They affirm that an inverse relation exists between musical tempo and the latency tolerance. Thus, when the delay increases, the tolerance or the allowed delay is smaller.

3.1.2.3 Timbre, pitch and intensity

The perception at the same time of two different sounds is strongly dependent on the characteristics of sound (timbre, pitch and intensity), that was declared at (Kleimola, 2006, p. 5) along with the idea of the previous paragraph.

3.1.3 Correlation

In these statements have been assumed that there is a relation between the physical delay, which is due to propagation, and the electrical delay introduced manually. However, is the physical delay introduced by the distance correlated with the electronic delay?

In (Schuett, 2002) an experiment was performed between two musicians who had to play a rhythm at different physical distances. After analysing the experiment, at a distance of 33m (100 ms) an incoordination in the rhythm was produced, such as in the experiments in which the delay is manipulated. Therefore, a correlation exists.

3.2 Optical delay measurement

As an optical delay, it will be understood in this document the time that an image spends to be shown on a screen after being captured by a camera.

This delay is composed of other delays like time-processing delay, transmission delay, and so on depending on which configuration or equipment is being used to transmit images from one end to the other.

3.2.1 Measuring methodology

The measuring has been done using an asynchronous method that due to real life, the camera is always shooting in an asynchronous way, and it is impossible to synchronise the shutter with the movements of an object.

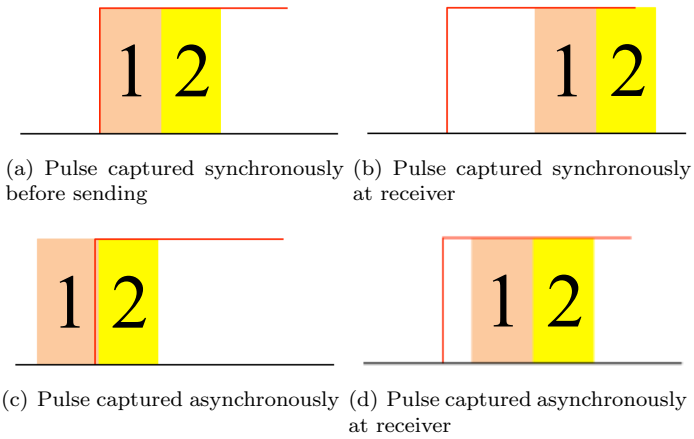
For the reasons previously stated, a range of possible delay values has been obtained. Seen from another point of view, one end is obtained when the shutter is synchronised with the event that wants to be recorded (figure 3.1(a)), getting the maximum value (figure 3.1(b)). In the other case, during a frame recording a tiny fraction of light is captured (figure 3.1(c)),. There can be a difference between these cases up to a frame depending on exposition time. This fact can be observed in figure 3.1. Afterwards, the delay will be treated and broken down thoroughly.

In order to have a completed control, the measurements were performed shooting with an exposition time equally to frame rate. Hence one frame is equivalent in duration to the shutter time.

This method consists in using a light-emitting circuit built with light emitting diodes (LEDs) and a function generator to get a periodic-squared signal of some hertzs that produces a flashing light when LEDs are excited. Then, this projected light signal is captured by a camcorder in an asynchronous way as it has been said before. This data is sent into different ways depending on the different scenarios to the projector, which emits the recorded light signal.

This light signal is transformed into electricity through a photodiode. An oscilloscope is the responsible for monitoring the signal that excited the LED and the signal captured by the photodiode. A clear sketch can be found in figure 3.2.

A comparison between them gives the delay measurement.



(a) Pulse captured synchronously before sending

(b) Pulse captured synchronously at receiver

(c) Pulse captured asynchronously

(d) Pulse captured asynchronously at receiver

Figure 3.1: Maximum and minimum delay values according to the instant when the pulse start is captured

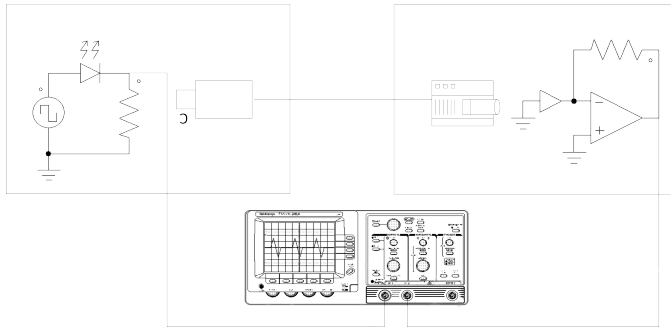


Figure 3.2: Equipment configuration to measure the optical delay

3.2.2 Signal Interpretation

Before starting to measure, an idea about what type of signal can be expected is necessary. Consequently, it must be borne in mind how the equipment that is being used distorts the signal.

As can be seen in the picture 3.3, it is appreciated how the blue signal (signal with peaks) seems not having any relation with the orange signal (rectangular pulse). However, the blue signal is the output of the rectangular orange periodic pulse through the system. To explain these changes, an analysis about how each particular device works took place.

3.2.2.1 Emitter circuit

It is the responsible for generating light signal, which corresponds to the electrical pulse transforming electricity into light.

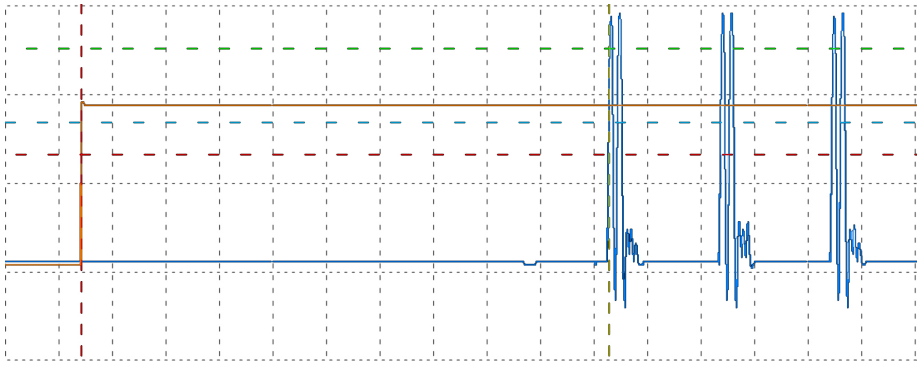


Figure 3.3: Incoming signal to the system (blue signal) and the delayed and modified outgoing signal (orange signal) by the system.

3.2.2.2 Photodiode

It transforms a light signal into electrical current, which can be monitored using an oscilloscope. To appreciate the impact that this device introduces on the generated signal, the photodiode was illuminated with the signal provided by the emitter circuit and no difference was noticed. A superimposed signal was obtained with a different level of intensity, thus the photodiode does not introduce any distortion to the signal.

3.2.2.3 Camcorder

It captures the light provided by the emitter circuit using a rolling shutter mechanism. This method has the drawback that different portions of the frame are exposed at different times than other portions. Hence, if a change is produced during the exposure, different effects like skew, wobble, or partial exposure, can appear in the frame. In fact, the image below (picture 3.4) was captured disclosing which sort of shutter mechanism is used by this camera.

3.2.2.4 Projector

The image is projected through a Digital Light Processing (DLP)[2] equipment, therefore a DLP projector uses micro mirrors that are switched to regulate an amount of light building a grayscale image. The colour is added modifying the white light provided by the lamp through the usage of a colour filter which tries to achieve the same effect that the Bayer grid performs, used to record coloured frames. Namely, the filter consists in a colour wheel composed of three basic Red, Green, Blue (RGB) colours. In special projectors, a great variety of colours can be introduced in this wheel to increase the number of combinations that the projector can show.

[2]DLP is a trademark owned by Texas Instruments

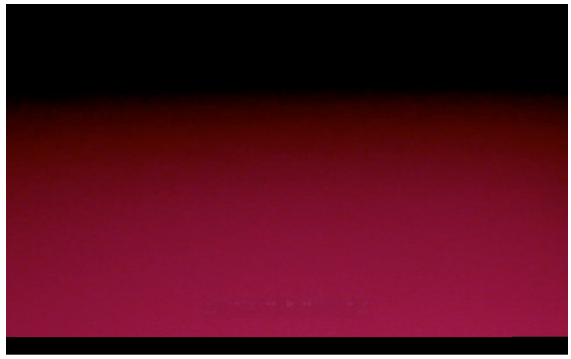


Figure 3.4: Frame in which the effect produced by the Rolling Shutter mechanism of the camera is appreciated

This colour formation modifies the output source in a considerable way.

In order to discover which are the features of the projector used, the following steps were taken:

The first step consisted in knowing which primary colours were used in its colour wheel. To overcome this aim, a photograph was taken moving the camera over the horizontal axis for a certain exposure time with the intention of obtaining all the changes made on that line for that time. Hence, it was concluded that the projector used a RGB colour wheel and it was able to obtain an estimation (figure 3.5) about how long the three colours are shown. As it has been seen below, this estimation is close to the values obtained using the second method presented.

To ensure that this values are coherent, an oscilloscope was used to look at the signal period presented by each colour. That was done by projecting a sequence of individuals colours. The three basic colours together with other colours which consisted of a combination of primary colours were observed (figure 3.6).

The period of the three colours does not correspond to a projector screening at 120Hz which can be explained since there are projectors that use a wheel with repeated colours. The frequency increases so that the different colours are shown.

In this case, this projector seems to use a RGBRGB wheel. Otherwise, the wheel is spinning twice per frame.

$$120\text{Hz} \rightarrow 8.33\text{ms} \qquad 240\text{Hz} \rightarrow 4.17\text{ms}$$

Now, all fit perfectly. Therefore, after doing these tests, it is clear why the pulse suffers these modifications. All changes happen as the projector shows only the colour captured by the camera according to the colour wheel. Consequently, a constant pulse does not appear anymore, instead of that, several peaks of colour that our eyes cannot distinguish are shown.

Weaknesses There is an important aspect that influences the measurements due to the measurement methodology and the projection working. The fact of using a

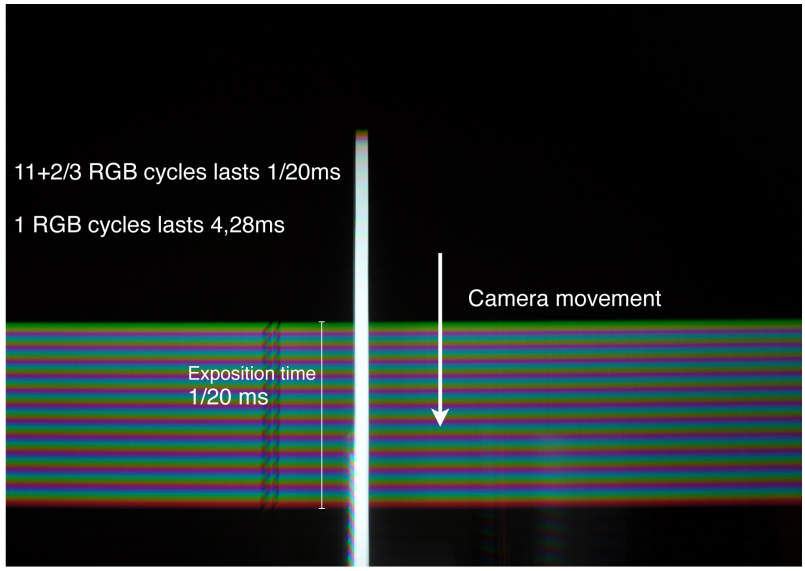


Figure 3.5: White colour generation with a RGB colour wheel

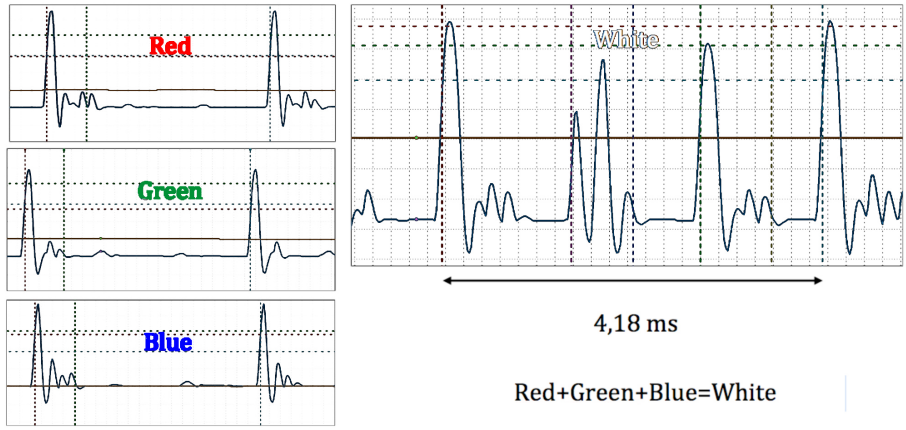


Figure 3.6: Colour waves

concrete colour and the colour wheel mechanism to screen it, produces a shifting of the value measured depending on which colour is chosen. For instance, there will be different measured values if the chosen colour is green instead of red because the colour wheel works sequentially. In fact, in figure 3.6 is depicted the three colours that make up the white colour, being able to appreciate that sequentiality. Consequently, the measurement must be referred to the same colour in order to characterize the equipment.

It is worth to say that this variation will be less than one RGB period, (see figure 3.6), and the tone colour depends on the LED's emitting light, camera and projector settings.

3.2.3 Values obtained

The test carried out consisted in measuring the optical delay that presented several configurations of systems. The equipment used is a sample of what was used in the musical rehearsal (see chapter 6). Besides this, other equipment provided by Uninett was used. That was an encoder and a decoder lended by Hitachi. This coding equipment has as its most important feature the ultra low delay that is introduced in the codification process. An overview of the main features of the used equipment can be consulted in Appendix A.

The values obtained have been taken for a long period of time due to the probability feature of this method.

Two different cameras has been used in this experiment because each device presents different type of interfaces.

As an example of this problem, both the encoder and decoder only accept Serial Digital Interfaces (SDIs) interfaces. To minimize this fact and to try to compare the acquired results with both cameras, converter boxes were used in several configurations. Nevertheless, a Digital Visual Interface - Digital (DVI-D) to SDI converter was not got. On account of these obtained results with one camera; an extrapolation of the entire system was done to compare performances.

The scenarios are presented from the simplest configuration to the complex ones.

3.2.3.1 Scenario 1

This is the simplest scenario which consists of a direct connexion between the end points. That measurement (table 3.1) was performed between the Toshiba camera (1080p@60Hz) and the projector via a DVI-D connexion (figure 3.7).

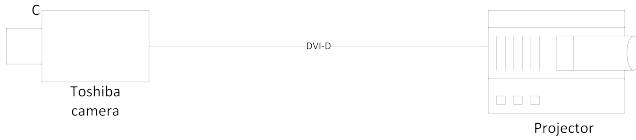


Figure 3.7: Physical configuration of Scenario 1

Duration of a frame $\frac{1}{60} \simeq 16.67$ms

Min value [ms]	Max value [ms]	Range observed [ms]
11.17	27.84	16.67

Table 3.1: Optical delay measurements of Scenario 1

3.2.3.2 Scenario 2

In this scenario, the same equipments have been used but changing the video resolution. Therefore, the Toshiba camera (1080i@60Hz) and the projector via a DVI-D connexion (table 3.2) has been connected. In principle, that projector is only able to manage progressive video, but this case was tried obtaining a reescale format (1090x540). This interlaced acquisition mode introduces at least 1 frame of delay, that is, the projector has to wait for 1 frame until receiving all lines of the frame for being projected progressively, thus a processing time also takes place. It would be equivalent to receive the images by the projector at 1080p@30fps but adding the interleaving time to make up the completed frame.

As it can be appreciated, the delay increased about 20ms (more than one frame) compared to the previous scenario.

That resolution change was done because of coding devices just can work at this new resolution.

Min value [ms]	Max value [ms]	Range observed [ms]
31.83	48.36	16.53

Table 3.2: Optical delay measurements of Scenario 2

3.2.3.3 Scenario 3

In this scenario, the measurements (table 3.3) were performed between the Panasonic camera (1080i@60Hz) and the projector via a HDMI connexion (figure 3.8) obtaining the same effect.

Figure 3.8: Physical configuration of Scenario 3a

Min value [ms]	Max value [ms]	Range observed [ms]
66.1	82.47	16.27

Table 3.3: Optical delay measurements of Scenario 3

These measurements have also been done using a SDI cable from the camera. A SDI to HDMI converter box was also employed (figure 3.9), obtaining pretty similar measurement. It can be concluded that the contribution made by this box for the delay is insignificant on this grade of precision.

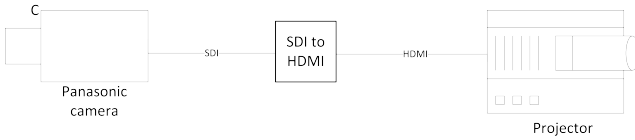


Figure 3.9: Physical configuration of Scenario 3 with converter box

3.2.3.4 Scenario 4a

New equipments have been included in these measurements (table 3.4) with the intention of reducing the amount of bandwidth needed in the communication.

Hence, the link between encoder and decoder requires less bandwidth. These measurements were performed between the Panasonic camera (1080i@60Hz) and the projector via a SDI connexion using a pair of encoder and decoder between them (figure 3.10). The encoder and decoder were connected directly through an Asynchronous Serial Interface (ASI).

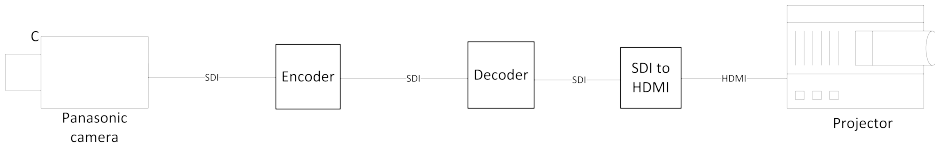


Figure 3.10: Physical configuration of Scenario 4a

Min value [ms]	Max value [ms]	Range observed [ms]
74.2	90	15.8

Table 3.4: Optical delay measurements of Scenario 4a

3.2.3.5 Scenario 4b

Measurements (table 3.5) were performed between the Panasonic camera (1080i@60Hz) and the projector via a HDMI connexion using a pair of encoder and decoder between them (figure 3.11). The encoder and decoder were connected directly through an ASI interface.

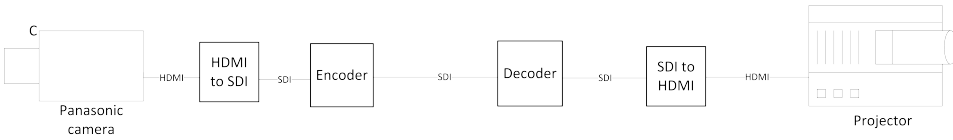


Figure 3.11: Physical configuration of Scenario 4b

Min value [ms]	Max value [ms]	Range observed [ms]
75.4	91	15.6

Table 3.5: Optical delay measurements of Scenario 4b

3.2.3.6 Scenario 5a

In this test, the encoder and decoder were connected via network interfaces using a switch. Consequently, this configuration could be used as a telepresence system using the current networks. In that case, the decoder can use a Forward Error Correction (FEC) function to increase the reliability of the system. If this option is enabled, an IP packet buffer stores the incoming packets. Hence, the processing time increases according to buffer size which can be configured in the setting menu. In that menu it is also displayed an estimation of the added time after setting this option.

These measurements (table 3.6) were performed between the Panasonic camera (1080i@60Hz) and the projector via a SDI connexion using a pair of encoder and decoder between them (figure 3.12). Furthermore, the FEC option was enabled which would increase the delay according to the configuration information about 8ms.

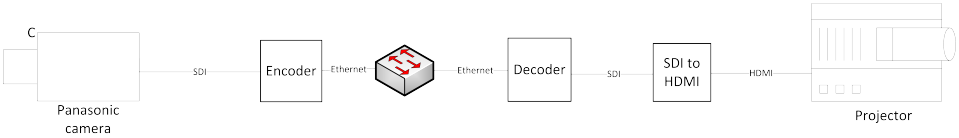


Figure 3.12: Physical configuration of Scenario 5a

Min value [ms]	Max value [ms]	Range observed [ms]
82.6	98.76	16.16

Table 3.6: Optical delay measurements of Scenario 5a

3.2.3.7 Scenario 5b

Measurements (table 3.7) were performed between the Panasonic camera (1080i@60Hz) and the projector via a HDMI connexion using a pair of encoder and decoder between them (figure 3.13). The encoder and decoder were connected via network interfaces using a switch and the same option settings mentioned above were set.

3.2.4 Extrapolation

After having obtained estimation values for the configurations presented above, it is possible to figure out the delay introduced for several used devices or just a few

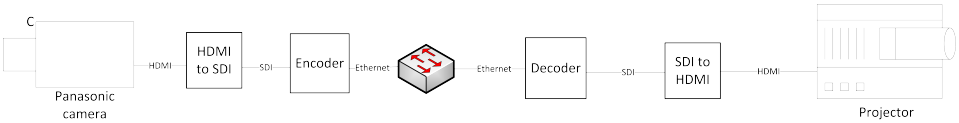


Figure 3.13: Physical configuration of Scenario 5b

Min value [ms]	Max value [ms]	Range observed [ms]
82.84	98.65	15.81

Table 3.7: Optical delay measurements of Scenario 5b

of them (figure 3.14). Through analysing Scenario 3 and 4 it can be obtained an estimation of delay contributed by encoder and decoder. It is about 8.3 ms. Doing the same technique with Scenario 4 and 5, it can be said that connecting these encoder and decoder via Ethernet interfaces when the FEC function is enabled (at the defined buffer size) increases the delay about 8 ms which is the same value that was estimated by the configuration. An overview of the obtained delay can be seen in figure 3.15.

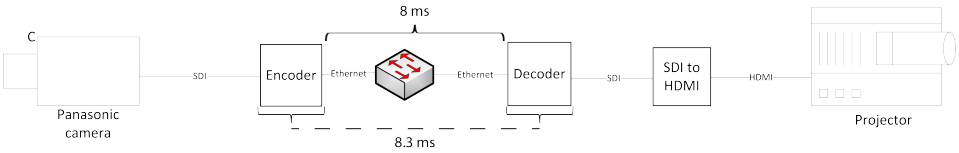


Figure 3.14: Delay introduced by equipment

Extrapolating these conclusions to the same scenarios' configuration but implementing the use of the Toshiba camera and a converter box to adapt the different interfaces will provoke an enhancement in performance in terms of delay, as it can be appreciated in figure 3.16.

3.2.5 Measurements in real networks

After performing several measurements in a controlled environment, the chance to repeat this test again through other networks arose.

This labour took place at headquarters of Uninett due to the fact that encoder and decoder boxes were being tested between this institution and the FCCN (Fundaçao para a Computação Científica Nacional). To this extent, the link was already established. However, because of this change of place the same equipment was not used. Hence the complete process of adding devices to figure out the delay introduced by each one was made.

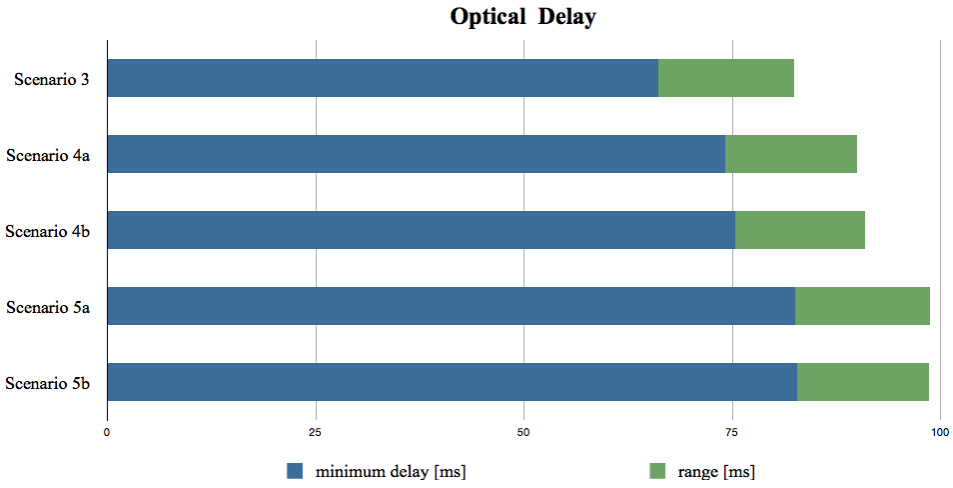


Figure 3.15: Optical delay measured using Panasonic camera

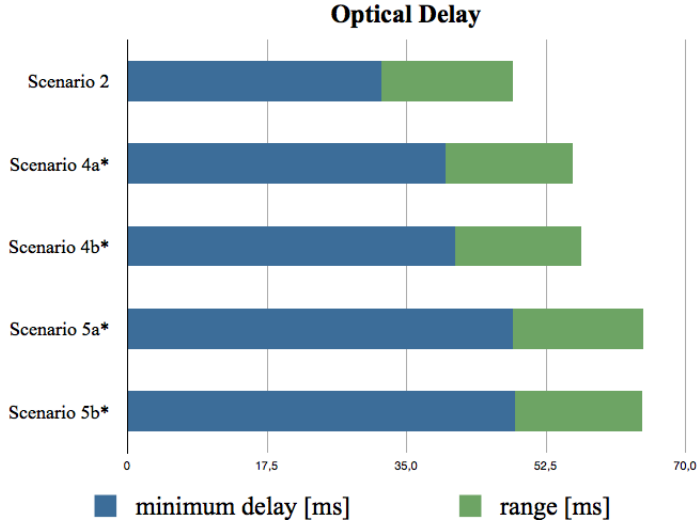


Figure 3.16: Optical delay extrapolation using Toshiba camera

Acronyms will be used to present the different configurations.

Camera	=	cam
Encoder	=	enc
Decoder	=	dec
Switch	=	sw
Router	=	rt
Server	=	server
Network	=	net
Projector	=	proj

3.2.5.1 Local measurements

These measurement scenarios are the same done in Sections 3.2.3.3, 3.2.3.4 and 3.2.3.6, but at this time the projector that was used does not store the frame when video streaming is recorded in interlaced way, it just forwards the frames. Moreover, that projector was not emitting a black background light when no lights were recorded, hence new peaks appeared in the oscilloscope display with regard to this new light signal. That drawback caused that the red peaks with low amplitude were harder to distinguish losing accuracy.

The configuration 3.2.3.4 in the scenario below will be used as a starting point to the process of adding the delay introduced by each equipment in order to compare it with the measured values.

For this scenario the obtained value using the new equipment was 49.2ms (minimum value) +16.2ms.

3.2.5.2 Reflection measurements

New scenarios were suggested as a way to check that the measurements and the method were being made properly. Thus, instead of sending out the packets locally via a switch, they were sent to a server which reflected the traffic using a small tool. This time, it was possible to remove the peaks caused by the background light, adjusting the luminance settings and other parameters of the projector. Moreover, ping time was measured to confirm that the measures had sense. Three scenarios were performed.

Uninett headquarter-NTNU-Uninett headquarter

The configuration performed is presented in 3.1 using the acronyms given above. The time measured using the methodology presented previously was 58.98ms (minimum value) +16.04ms which is fairly similar to the value obtained (equation 3.2) when adding the calculated time introduced by the preceding iterations.

$$\text{cam} \rightarrow \text{enc} \rightarrow \underbrace{\text{sw} \rightarrow 4 \cdot \text{rt} \rightarrow \text{server} \rightarrow 4 \cdot \text{rt} \rightarrow \text{sw}}_{\text{ping time=0.89ms}} \rightarrow \text{dec} \rightarrow \text{proj} \tag{3.1}$$

$$49.2 + 0.89 + \underbrace{8}_{\text{Buffer time}} = 58.09\text{ms} \tag{3.2}$$

Uninett headquarter-Oslo-Uninett headquarter

In this new scenario (sketch 3.3), the measured time was 67.3ms (minimum value) +16.3ms. In equation 3.4 it is possible to appreciate how this time, the values also

match themselves.

$$\text{cam} \rightarrow \text{enc} \rightarrow \underbrace{\text{sw} \rightarrow 6 \cdot \text{rt} \rightarrow \text{server} \rightarrow 6 \cdot \text{rt} \rightarrow \text{sw}}_{\text{ping time=9.16ms}} \rightarrow \text{dec} \rightarrow \text{proj} \tag{3.3}$$

$$49.2 + 9.16 + \underbrace{8}_{\text{Buffer time}} = 66.36\text{ms} \tag{3.4}$$

Uninett headquarter-Svalbard-Uninett headquarter

Finally, the same process (sketch 3.5) was repeated to a server, which was placed far away obtaining similar conclusions. The time obtained was 91ms (minimum value) +16.4ms, similar to the value of equation 3.6.

$$\text{cam} \rightarrow \text{enc} \rightarrow \underbrace{\text{sw} \rightarrow 14 \cdot \text{rt} \rightarrow \text{server} \rightarrow 14 \cdot \text{rt} \rightarrow \text{sw}}_{\text{ping time=33.5ms}} \rightarrow \text{dec} \rightarrow \text{proj} \tag{3.5}$$

$$49.2 + 33.5 + \underbrace{8}_{\text{Buffer time}} = 90.7\text{ms} \tag{3.6}$$

As it has been observed, the predicted values by the ping tool (these values are an average of several measurements) fit perfectly with the transmission time of sent packets.

In these cases, the packets travelled through the Uninett network. In the section below, the measurements were done through an heterogeneous network.

3.2.5.3 Round trip measurements

This time, the measurements were taken with the collaboration of the FCCN (Fundaçao para a Computação Científica Nacional).

Direct connection

In this presented configuration (sketch 3.7), the encoder and decoder (the same coding devices were used) at Portugal were connected directly using the network capabilities. The FEC function was enabled adding more delay. In fact, the extra delay configured and estimated by the device was 133ms. Therefore, this value is used in equation 3.8 to verify whether the measurements (270ms) fit with the estimation.

$$\begin{aligned} \text{cam} \rightarrow \text{enc} \rightarrow \underbrace{\text{sw} \rightarrow \text{net} \rightarrow \text{sw}}_{\text{ping time=74.5ms}} \rightarrow \text{dec} \rightarrow \text{enc} \\ \rightarrow \overbrace{\text{sw} \rightarrow \text{net} \rightarrow \text{sw}} \rightarrow \text{proj} \end{aligned} \tag{3.7}$$

$$49.2 + 74.5 + \underbrace{8}_{\text{dec-enc}} + \underbrace{133}_{\text{Buffer time at FCC}} + \underbrace{8}_{\text{Buffer time}} = 272.7\text{ms} \tag{3.8}$$

Optical connection

In this new case, at FCCN (Fundaçao para a Computação Científica Nacional) the signal was projected on a screen and it was recorded again to be sent back into the system (sketch 3.9). Consequently, the asynchronous problem mentioned at beginning of this chapter appeared again. In fact, the measured obtained was $351 + 31.8$ms, almost 2 frames of variation.

$$\begin{aligned} &\text{cam} \rightarrow \text{enc} \rightarrow \underbrace{\text{sw} \rightarrow \text{net} \rightarrow \text{sw}}_{\text{ping time=74.5ms}} \rightarrow \text{dec} \\ &\rightarrow \text{proj} \rightarrow \text{cam} \rightarrow \text{enc} \rightarrow \overbrace{\text{sw} \rightarrow \text{net} \rightarrow \text{sw}} \rightarrow \text{proj} \end{aligned} \tag{3.9}$$

$$49.2 + 76.5 + \underbrace{8}_{\text{dec-enc}} + \underbrace{133}_{\text{Buffer time at FCC}} + \underbrace{8}_{\text{Buffer time}} + x = 351\text{ms} \tag{3.10}$$

The introduced delay by all devices was impossible to measure since part of them were in Portugal. Nevertheless, in the equation 3.10 being x the delay introduced by the camera and the projector, an estimation was obtained: $351 - 274.7 = 76.3$ms which resembles values obtained in scenario presented in 3.2.3.3.

3.2.6 Improvements in the measurement method

One way of increasing the accuracy of this methodology would be to automate the process of measurement. As it has been mentioned above, two extreme values determine the range of possible delay values. Thus, instead of acquiring a large amount of values that are inside this range, the extreme values can be obtained.

3.2.6.1 Obtaining the right end value or maximum delay

To get this value, the signal that wants to be recorded (light signal) has to be synchronised with the shutter of the camera. Nowadays, there are a lot of cameras that incorporate an interface for synchronization. Furthermore, most function generators can send out a synchronization signal at the same time than they are generating the output signal. Besides this, some of them are also capable to receive a synchronization signal to generate the output according to that signal.

The only possible problem that can appear is that these signals can be of different types. For example, the function generators used to work with Transistor-transistor-logic (TTL) signal whereas modern camcorders uses tri-level signals. Hence, the layout presented in figure 3.17 matches with both configurations. The difference stems from which device the sync is provided.

If the sync is provided by the function generator, the 'Signal sync' module will focus on the adaptation of the signal received into one which electrical values are compatible with the camera sync signal.

In the other case, the 'Signal sync' module would provide the right signal to both devices respectively.

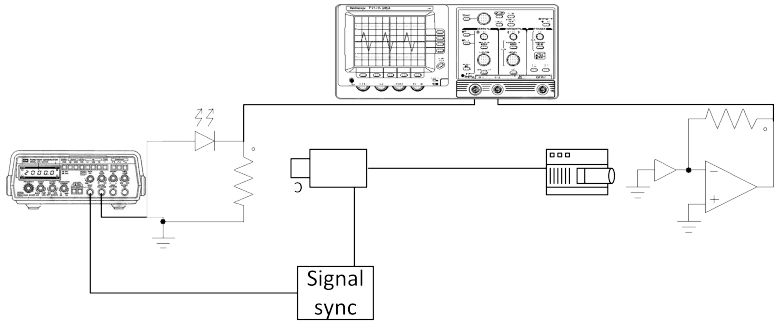


Figure 3.17: Syncrhonization cirtuit

3.2.6.2 Obtaining the left end value or minimum delay

This value or the signal related with it, appears when the shutter time still captures a minimum quantity of light before starting a new frame. Therefore, the proposed methodology to get this value is not as accurate as the previous method. It consists in performing the measurement as the previous measurements have been done, that is, asynchronously. The difference lies in how the value will be read. Instead of acquiring them manually, the envelope mode of an advanced oscilloscope would be used.

This mode combines the minimum and maximum samples from multiple acquisitions to form a waveform that shows minimum and maximum changes over time.

Thus, the measurements would be carried out for a long period of time without human interaction. The extreme value will be obtained automatically.

3.3 Methods

When getting information about a system in which exists a human interaction, one of the first issues to overcome is how to get this data.

However, it is worth to mention that it is possible to achieve two types of practical information, objective information or subjective information. Usually, the greatest challenge to get objective information is to find the right procedure, which means, how to measure the target. If subjective information is pretended to be acquired, the issue stems from the validation of truthfulness information provided by personal opinions.

This chapter is focused on subjective data. Nonetheless, some methods to get objective information have already been presented in previous sections regarding the delay.

3.3.1 Surveys

As everyone must know, a survey is a set of questions which an individual is asked to respond. It is based on a sample of the population. Whereby, it can be a double-edged sword. One of the benefits of a survey is the facility to get enough responses. In contrast, the validity of this data must be thoroughly analysed.

In some cases, surveys are used because they are the easiest method. But this is not our case.

Target

Once it is decided to prepare a survey, the population of interest or the object have to be set. Sometimes a survey is needed to get a complete vision of the population but in other occasions, it is preferable to get the values to group them according to their characteristics. In the presented case, the participants would be the people who have been testing the system. According to these features, questions were focused on them.

Writing surveys

The achievement consists in composing a set of well-written questions, avoiding biased issues. Different type of questions can be distinguished:

Open-ended questions, which are useful to get more complete answers. Mainly, these sort of questions produces more tedious data to analyse. The individual interviewed can answer responds that are not expected, being out of what was asked. Furthermore, a good question is difficult to prepare.

Close-ended questions, can be a set of different ordered questions. These questions can be set according to degrees of agreement, or simply multiple-choice questions in which you have to choose one option or more than one possible answer.

Structure

The structure of the survey must be clear. It must explain in a compressive way how the respondent has to fulfil the questions at the beginning. The related questions must be placed together. It must be taken into account that the respondent does not have to be overwhelmed for the survey. Thus, the form does not need to be too long.

Testing

Another important point is testing. When survey has been developed, it must be tested before being presented to targets. All questions must be checked deeply to accomplish the requirements above.

3.3.2 Structured interviews

An structured interview is just an interview where the person is asking another person a batch of well-selected ordered questions.

Benefits

The goal of this method is to provide an structure to the interview. In this way, all questions will be asked to the people in the same order. Furthermore, a criteria to rank these questions can be defined and fulfilled once that question has been answered. Moreover, the interviewer can explain and clarify to the respondent all sort of misunderstanding or confusing questions. Among other benefits, the fact of having a predefined structure allows to compare the answers of all participants in a clearly way without any ambiguity. Besides this, a comparison can also be made between surveys that have been filled at different times.

Regarding the reliability of this method is quite high, but it depends on the participants.

Drawbacks

Unlike a survey, making this sort of activity is a time-consuming process not only by the fact that it requires a thorough prior preparation to have a good choice of questions but because the interviewer has to be present in each interview.

Concerning the questions, they are fixed. The interviewer cannot change the content once the process has started. Hence, the fact of asking a new question to obtain extra information is not allowed.

Moreover, the answers are usually characterized by a lack of details.

Chapter 4

Reaction Time

4.1 Introduction

Human beings react by nature to any stimulus causing them sensations. These feelings often produce an instantaneous response as a defence mechanism. Otherwise, feelings can also be analysed by the individual generating a proper response to these stimulus.

The stimuli can be of any nature (visual, audible...). However, the reaction is different depending on which has been perceived.

One of the aims of a telepresence system is to try to get the same perception or reaction that a person experiences in a common environment. Notice that in this case, the stimuli is produced by the system. Does it greatly modify the user's response?

4.2 Overview of reaction time

Reaction time is defined as the time between the onset of a stimulus and the beginning of an overt response Coren et al. (1984). Different paradigms have been proposed to measure the Reaction Time (RT) and they can be grouped into several categories.

Simple

It is the time that a subject needs to respond to a single stimulus. For instance, a subject must emit a sound when a red frame is shown on a video.

Recognition

In this case, several stimuli are presented and the response must be made only when one type appears. For example, various coloured frames can appear in a video and the individual needs to react when the red frames are shown.

Choice

In this paradigm, the individual being tested will respond in a different way to different stimuli. For instance, the individual will emit a different sound depending on the colour which has been observed in the video stream.

It is worth to say that the reaction time is influenced by numerous factors. Thus, a lot of researches (Kosinki, 2010) took place to try to figure out how these factors increase or decrease the duration of the reaction time. Some of them are mentioned further on, and a discussion about how they can affect the measurements will take place.

4.2.1 Influencing reaction time

Age

As people get older, the response time to a stimulus begins to grow. In addition, there are three different stages. In the first one, the ages included from childhood to 20s, the lowest reaction time interval is detected. During the period comprised from the 20s to the 50s, the RT grows slowly, reaching the highest values at 70s or more. Researchers have proposed several explanations to this fact. They claim that is not just by body's debilitation because of the age. Perhaps, maturity influences analysing people's acts deeply.

Gender

Males and females have different reaction times. In fact, male's reaction time is shorter than female's.

Practice

The fact of practising an action for many time helps the reaction time to decrease. In a scenario where one person is doing an action for the first time, this person's reaction will take a long time whereas the person who had been practising it many times will need less time.

4.2.1.1 Other factors

As it has been said above, there are many factors that could affect the reaction time. For example, distraction, fatigue or either eating some substances can modify our reaction time. Others like intelligence or personality type do not depend on each person's decisions.

4.3 Frame rate and reaction time

The frame rate is one of the parameters that characterizes a video stream. It indicates the number of frames (unique images) that are shown per second. The eye has a visual sensitivity. It can distinguish different images that are shown below

a specific frame rate that depends on each individual. Moreover, what we want to discover is whether an increase of the frame rate will provoke a stimulus-response time improvement.

Furthermore, we would like to try to figure out whether there is a higher threshold at which the increase in frame rate does not affect the stimulus response. Therefore, this research has been focused on light stimuli, using the simple reaction time paradigm.

The experiment consisted in playing a video in which an event happens in an instant of time. This video was recorded at different frame rates. A thoroughly explanation will be done below.

4.3.1 Methods to measure reaction time

The methodology proposed for the test was based on measuring the time passed between a visual stimulus and the corresponding response. In this case, that response was a sound emitted using the voice.

Therefore, to achieve this goal, different approximations were proposed as a way to decide which equipment use and how to obtain the results the more accurate as possible.

To make this choice, a coloured frame change was used as a stimulus signal. That fact eliminates the influence of frame rate on video reproduction.

4.3.1.1 First approximation

The idea consisted in using a projector to show the visual stimulus and a microphone to mitigate the propagation time of acoustic waves. To measure the spent time between the first event and the vocal response, an oscilloscope was proposed. However, a reference signal that specifies the instant in which the event has happened into the oscilloscope was also needed.

In order to get this signal, an accurate video-editing program was used. This program could insert an audio signal into the right place of the video. Afterwards, this signal was sent out to the oscilloscope defining the instant when the stimulus occurred. The fact of using an audio signal, which would not be listened by the individual under test, stems from the fact that it will be easy to monitor this signal using an oscilloscope. This signal will be send directly, without any conversion.

In principle, this method seems not to have any problem but how it works is analysed below.

Drawbacks The fact of using a projector to present the image to the person who is doing the experiment produces an incoordination with the reference signal. The reference signal (audio signal) was directly connected to the oscilloscope whereas the stimulus signal was sent out to the projector. Besides this, the projector needs a processing time to screen the images.

In general, this would not be a problem owing to the fact that the introduced delay between both signals can be figured out and it would be permanently constant.

Measuring the lag between signals In order to do this task, a video was used as a stimulus signal. This video was composed of black frames, but in some instants of time red frames were inserted between a pair of black ones. Therefore, a black video sequence which included red frames in some moments was created. The red frames were the events, and whereby, the reference signal was located in the correct place when the video was changing into red.

It was surprising that the lag observed was not constant. This phenomenon occurred due to the operating system process scheduler. The video was played using a computer, which is a machine that can execute several processes or data flows with different priorities, and the video and audio streams are managed by different flows. Thus, it is possible to increase the priority of a video player executed in a computer, but the operating system processes always have a higher priority than a user process. It is worth to mention that the video was played and sent out to the projector as well as the audio, which was sent to the projector using another interface.

Consequently, if the video was played, the priority obtained was not the highest due to the fact that this process was put off the queues. When that process was in execution was inserted in the queue of the running process.

What was observed when running this video on a computer with Mac OSX was that the lag was varying very slow as the video was progressing. The audio signal was connected via a mini-jack cable. The video used a Mini DisplayPort to DVI-D adapter and a DVI-D cable. The border values can be seen at figure 4.1, giving us around 17ms of variation.

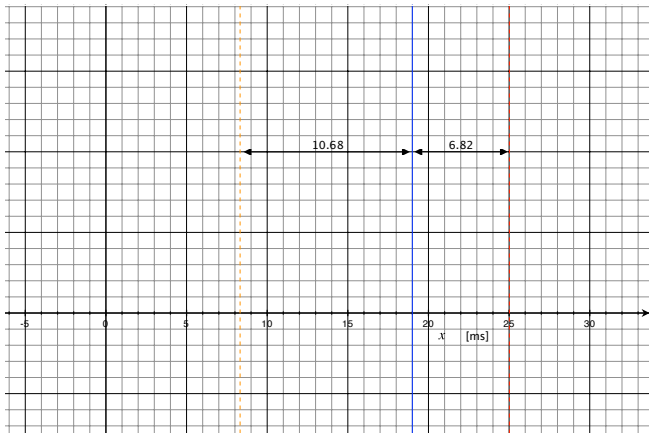


Figure 4.1: Lag between audio and video signal on a Mac OS X computer

Nevertheless, if the video was paused, the process was put off the queue, and the priority assigned once the video at that time was in reproduction again was apparently random.

Once the video was played again, the signal monitored in the oscilloscope seemed to be in a Round-Robin queue owing to the fact that the lag was varying in this

way. That is, in an instant of time the signal was approaching to one end, and then the lag signal went to the other end starting again.

In favour of avoiding this problem and not to lose accuracy in these measurements, the same experiment was repeated using a PlayStation 3 (PS3) as a video player. In this case, the output video signal was connected to the computer using an HDMI cable and the audio signal was sent out to the oscilloscope using a RCA cable. The same process was repeated using a ordinary DVD player. However, the same drawback appeared again.

4.3.1.2 Second approximation

As an accurate result wants to be gotten, the priority or the schedule discipline of the operating system would have to be modified. The video would have to be run in real time in order to avoid the lag between the stimulus signal and the reference signal.

With the purpose of avoiding doing this, it was proposed to use the stimulus video as a reference signal. Thus, at that time one signal was achieved instead of two, avoiding this problem.

This method consists in modifying the video stream using some pixels as a reference signal. To acquire the video signal, a photodiode was used again. The photodiode is sensitive to the entire visual spectrum except for the black colour, that means, there is no captured light with this colour. Consequently, a small area of the video stream would have to be always black except when the stimulus occurs. In this case, this area was coloured. The red colour was chosen because the photodiode used in this experiment had the highest sensitivity peak in this red colour.

Drawback The user experience is being interfered by using this method. In fact, the event recognition can be influenced by this signal, which can be a clue. It is worth to mention that this coloured signal takes place for one frame, being perceptible to the human eye when the frame rate is not high. In order to mitigate this fact, this part of the frame has not to be shown to the individual under test.

Measuring the lag between signals As it is sketched in the figure 4.2, it is possible to perform a background projection, hiding the area that does not want to be shown. In fact, in the projection's frontal side of the screen the photodiode was placed, and in the other side, the individual under test was not able to see the area used to the reference signal.

In order to acquire the value, only the two signals have to be compared using the reference signal as a trigger.

4.3.2 Measuring reaction time

As it is obvious, the second method was chosen and different experiments were proposed.

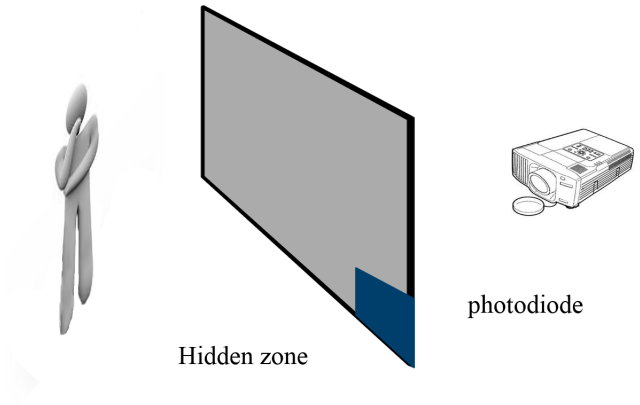


Figure 4.2: Scenario proposed to measure the response time using the second approximation

4.3.2.1 Phonetic studio

Before doing the experiment it was necessary to perform a small phonetic studio to obtain which sound would be better distinguished. The aim was to get a wave with a steep slope and a high intensity visualized in the oscilloscope.

For instance, the pronunciation of vowels produces for a short period of time a very poor intensity. This intensity increases as the amount of air is going out of the mouth. Hence, several combinations were tried, obtaining the best result when the phoneme /pa/ was pronounced.

4.3.3 Performed measurements

The results obtained cannot be generalized because of the multiple factors involved in such measurements, which have been previously explained. Furthermore, only around a hundred samples were collected in each experiment, involving four people. Participants had no relation with the project except for the author of this thesis. They signed a document (Appendix G) that allows the data publication. A larger study should be done to validate this data with a more global framework. However, they can be taken as an hypothesis. It will be applied for the following tests too.

Scenario A This test was made to verify the method used and to get a reference value for posterior tests. The stimulus video signal is the video which was described in the first paragraph of Section 4.3.1.1. Consequently, a classic visual stimulus response measurements were performed to the participants.

When analysing the collected data, it could be observed that a lot of researches have been done using simple descriptive statistics. These methods are focused on reporting a central tendency parameter like the mean, and a dispersion parameter like the standard deviation.

However, recent researchers indicate that those methods are unsuitable to analyse Reaction Time (RT) data. In Whelan (2008), performing an analysis of the data's statistical distribution is proposed as a method to analyse response time. The ex-Gaussian distribution used, which has been demonstrated, fits perfectly with the RT analysis. For this reason, and using the Lacouture and Cousineau (2008) developed functions with MATLAB, the figure 4.3 was plotted. In this graphic, the histogram has been depicted regarding the acquired data and two functions that try to fit to this. As it can be appreciated, the best fit function corresponds to the ex-Gaussian distribution.

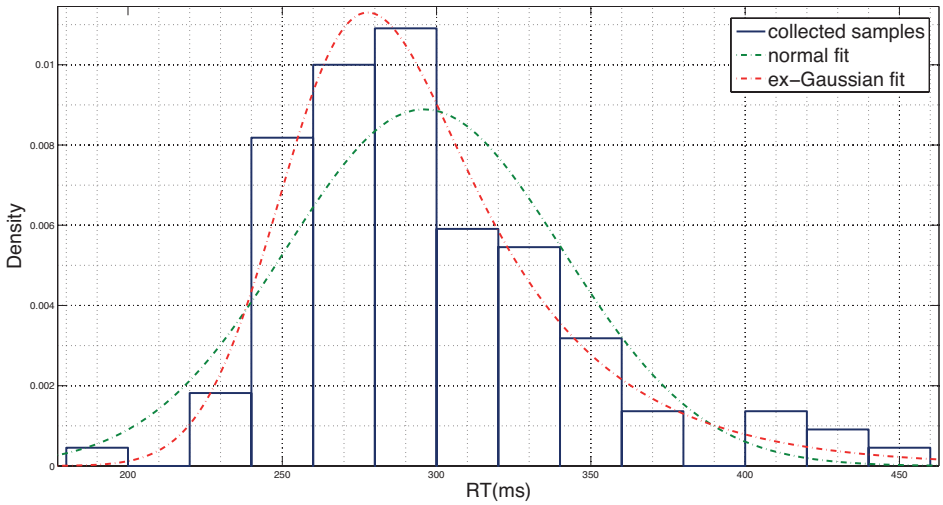


Figure 4.3: Probability density function (PDF) of RT samples acquired in scenario A compared to Ex-Gaussian and Normal distributions

This function is the result of the convolution of a Gaussian (normal) function and an exponential function. It is characterized by three parameters, $\mu = 256.3175$, $\sigma = 22.1997$ and $\tau = 39.6093$. The former refer to the Gaussian component corresponding to the mean and standard deviation, respectively. The last one is related with the exponential component, describing in one parameter the mean and the standard deviation.

Therefore, the mean (equation 4.1) and the standard deviation (equation 4.2) of this new function is composed of additive functions that will be as follows:

$$\overline{x} = \mu + \tau = 295.9268 \tag{4.1}$$

$$s = \sqrt{\sigma^2 + \tau^2} = 45.4062 \tag{4.2}$$

Scenario B In this scenario, the focus was to discern the relationship between frame rate and stimulus response. The stimulus that was presented in the recorded videos at different frame rates was a small box falling down. A red line was inserted in the middle of all frames giving us the instant when the individual under test had to emit the sound. Expressed in a different way, when the box was at the same height that the red line, that person reacted to this fact. In this video, the participant had enough time to predict the trajectory of the decline and this cooperator also had the enough time to emit the response (the sound) at right instant of time.

Tests consisted of three video at 30, 60 and 120fps. The original video was recorded at 60fps and the other two were obtained from the original clip. Therefore, to obtain the 30fps sequence, half of the original video frames were dropped. In contrast to the 120fps case, interpolation algorithms were used to predict and construct the missing frames.

Once the test finished, participants were asked whether they noticed any difference among the three videos. The answer was unanimous, everybody said that they had the feeling that the box seemed to fall faster in the recorded video at 30fps.

This event can be attributed to the difference in the frame rate. Regarding the other two videos (60 to 120fps), no discernible difference was found. According to this similitude, there are two possible options. The first chance is that 60fps is sufficient for the video shown. The other circumstance will be that the 120fps video was obtained by interpolation. This interpolation mechanism does not represent the quality of a video recorded at this rate.

It is worth to mention that the reaction time in the first samples collected tend to be high. However, when the individual is able to predict the trajectory, the obtained values are close to whenever the event takes place, even emitting the sound some milliseconds before the event occurs. These negative values were taken out of the samples.

The parameters that characterize the probabilistic functions can be seen in table 4.1 . These functions fit with the collected data. Furthermore, the corresponding graphs to these measurements are presented in Appendix B.

	μ	σ	τ	mean	standard deviation
30 fps	44.7907 ms	34.3411 ms	27.7984 ms	72.5891 ms	44.1822 ms
60 fps	7.2077 ms	5.5370 ms	49.7331 ms	56.9408 ms	50.0404 ms
120 fps	31.3429 ms	23.9628 ms	26.7382 ms	58.0810 ms	35.9046 ms

Table 4.1: Characteristic parameters of probability distributions of Scenario B

Paying attention to the mean's column, it can be appreciated that there is a signifier difference between the video recorded at 30fps and the other two sequences. It was noticed during the experiment that the time necessary to predict when

the event took place was longer in the 30fps video. Therefore, more values were captured with a longer reaction time.

Concerning the 60fps and 120fps videos, there is no great difference to the mean parameter. However, this discrepancy is considerable regarding the other parameters. This dissimilarity is obvious in the fit functions as it can be observed in figure B.4. Nonetheless, the histogram of acquired samples (figure B.5) is almost identical, having a divergence from 20 to 40ms. If this difference did not exist, the fit functions would be more analogous. Therefore, there is no big difference between both videos.

Scenario C This test is based on the previous scenario, the only particularity was that this time it was fairly harder to predict the trajectory. In order to avoid this problem, the viewing window was narrower.

Most of participants reported that they did not notice any difference among the videos. This fact can be justified because the event occurred very fast. One of the cooperators said that it was a bit easier in the video of 120fps.

The parameters that characterize the probabilistic functions and fit with the collected data are presented in table 4.2 (see figures B.7, B.8 and B.9).

	μ	σ	τ	mean	standard deviation
30 fps	152.4530 ms	21.0466 ms	28.7821 ms	181.2351 ms	35.6563 ms
60 fps	156.0139 ms	27.9649 ms	34.1613 ms	190.1752 ms	44.1478 ms
120 fps	136.2413 ms	28.9769 ms	39.6618 ms	175.9031 ms	49.1194 ms

Table 4.2: characteristic parameters of probability distributions of Scenario C

In this Scenario C, the mean values are very similar. Moreover, the fit function parameters as can be seen in figure B.10 are akin too. The same phenomenon occurs with the samples (figure B.11). Likewise, it is possible to conclude that the speed with which the event happens causes that the reaction time tends to a stimulus in which there is no possible prediction regardless of the frame rate.

Chapter 5

Space perception

One of the problems of the current telepresence system is the lack of space perception. That is, the depth sensation or ability to be able to make out different distances among objects.

These systems, as interface to communicate with the other end, have a flat screen in which 2D video is shown. In some cases, the 2D video eliminates the space perception. In others, this perception is decimated due to the fact that monocular cues also aid.

This is not a problem either for a videoconference or when a simple communication wants to be established with the other end. The problem appears when the aim of the communication is wider than this phenomenon. For instance, when a person needs to interact with the other end pointing objects.

In fact, one of the problems that some musicians reported in the research (Konstantas et al., 1999) carried out between the University of Geneva and the German National Research Center for Information Technology was the impossibility of gestural designation.

As a measure to try to solve or at least mitigate this lack of space perception, some methods were proposed to perform among ends to conduct some test.

5.1 Stereoscopic 3D

Stereoscopic 3D consists in using two-dimensional images of the same scene, but they are taken from different points of view to create the illusion of three-dimensional depth tricking the brain into merging these into one. Therefore, the 3D effect is achieved owing to image disparity that appears in the screen, which produces a fairly similar effect at the retina. Moreover, human vision for measuring the depth uses more cues. ie:

Monocular cues

- Focus
- Perspective

- Occlusion
- Lighting and shading
- Colour intensity and contrast
- Relative movement
- Vergence

Binocular cues

- Stereopsis

A wider explanation of these cues can be consulted in Appendix C.

5.2 Showing 3D objects

As it is said before, two images are projected onto a display. The method used to distinguish these images corresponding to each eye in stereoscopic 3D consists in using a pair of special glasses. In order to get this purpose, there are several artefacts (shutter LCD, polarization, filtering) but all of them are based on the same idea. Each eye has to see the channel signal corresponding to it.

But in all of them, the images will be slightly horizontally displaced. After that, the brain will have to merge these objects into one. The only way to achieve it, consists in assuming that the object is either in front or behind the screen plane.

Three cases can be distinguished (Figure 5.1):

In 5.1(a) the object appears behind the screen because the brain interprets that the vanishing point converges behind the screen. In 5.1(b) the convergence takes place onto the screen, like in a 2D picture. The last picture (5.1(c)) shows the object in front of the screen, where the brain reconstructs the image.

These situations will depend on how the video is recorded and how the cameras (eyes) are situated.

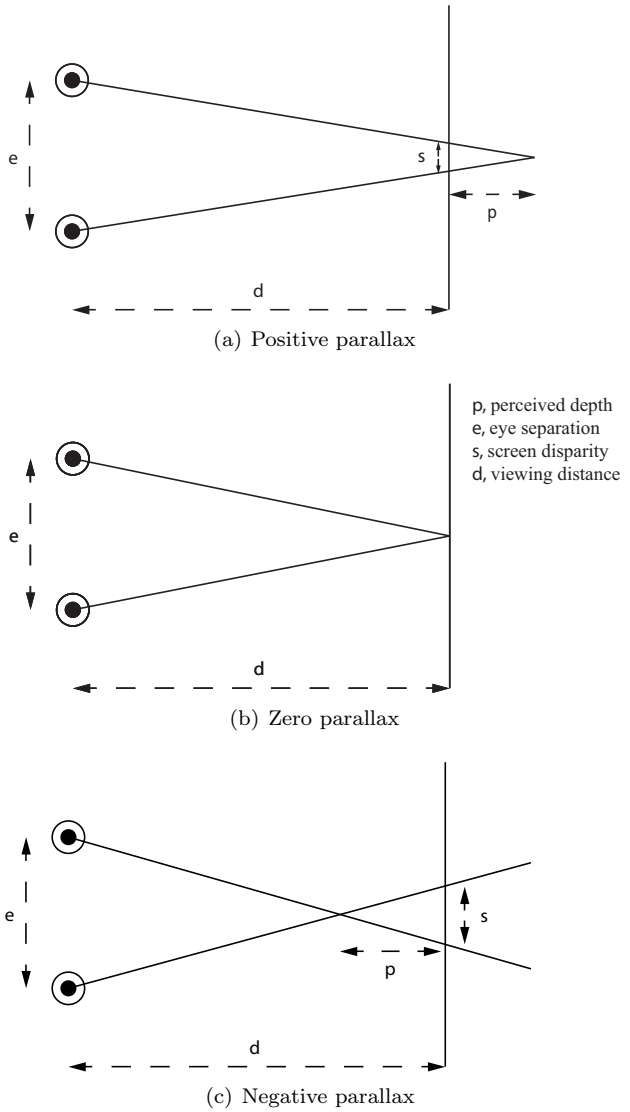
5.2.1 3D Shooting

Depending on how the scene is recorded, the different distances perceived will be different. In other words, image disparity is modified.

The standard 3D shot consists in placing the cameras separated at a distance about 65mm, which is the average distance between human beings' eyes, parallel to one another. In this case, the convergence point is situated at infinity, causing that the whole scene appears in front of the screen.

Moreover, if the distance between the two cameras is increased, the perceived depth will increase as well. Additionally, the distance among objects will increase as well if these items are located in different planes.

When the angle between the two cameras is altered, making it an acute angle, the vanishing point will not be the infinity. Hence, the perceived 3D image can appear behind the screen.



(a) Positive parallax

(b) Zero parallax

(c) Negative parallax

Figure 5.1: Types of parallax effect

Furthermore, the different shooting possibilities in a video recording must be taken into account to decide which 3D effect or sensations want to be gotten.

5.2.2 Variation in perception

Pixels Any image is made up with pixels which are a small area representing one colour. In the ideal case, a pixel should be a point but they actually have a

considerable area. As a result, they cannot represent a point with depth, instead of that, they constitute a volume of perceived depth.

Hence, there is a threshold of depth. Depths differences among points that are under the depth limit cannot be represented.

Screen disparity Image disparity is constant for a given stereo pair, but screen disparity is not fixed. It depends on display features.

$$\text{On-screen disparity} = \text{image disparity} * \text{pixel size}$$

However, if the geometric model of perceived depth used in 5.1 is taken and analyzed, new variables appear.

In the positive parallax (figure 5.1(a)) the following perceived depth is obtained.

$$p = \frac{d}{\left(\frac{e}{|s|}\right) - 1} \tag{5.1}$$

And (figure 5.1(c)) is obtained for the negative parallax

$$p = \frac{d}{\left(\frac{e}{|s|}\right) + 1} \tag{5.2}$$

Viewing distance (d) Perceived depth (p) is directly proportional to the viewing distance (see equations 5.1 and 5.2). Consequently, the individual who s closer to the screen perceives less depth that the person who is placed further away.

Eye separation (s) Perceived depth is inversely proportional to individual's eye separation (see equations 5.1 and 5.2). ie. the person who has the eyes closer perceives more depth.

For the reasons stated above, all these parameters and some others must be taken into account to perform an accurate 3D representation. On the contrary, a fake representation will be obtained. This representation, perhaps, is enough for some applications. Therefore, a tracking of these parameters must be done in applications where the depth is critical.

5.2.3 Overview 3D formats

Most 3D products are intended to create or to manipulate the content for digital entertainment. Hence, all products take into account the capabilities of end-devices like 3D TVs. Most of them are characterized by being able to display 3D content in a Frame Sequential way. However, they can support several 3D formats as input. The conversion takes place at the same time that the content is being displayed.

5.2.3.1 HDMI 1.4 capable 3D formats

HDMI 1.4 is the *de facto* standard that defines how to transmit 3D content through this interface as well as the physical requirements and improvements made.

Among the many supported formats, the most used are:

Side by Side 3D Format This format is Frame Compatible, which means that this structure can be used in HDMI 1.3 compatible devices.

Each frame of side-by-side 3D consists in an horizontally scaled combination of the frames for the left and the right eye. That means, a single frame that actually contains sub-frames for both left and right eye.

In the Frame Compatible format or "Side-by-Side Half" each sub-frame is sampled in such a way that they have only half of the horizontal resolution of a true HD frame (for example 960x1080 instead of 1920x1080 for 1080p content), . This mechanism allows each frame to have the same size as regular 2D HD, making it easier to transmit and compatible with HDMI 1.3 equipment.

It is possible to stack also full HD frames side-by-side for full HD 3D, but this format is not mandatory as part of the HDMI 1.4 specification.

Top-and-bottom or Over-Under 3D format It is also Frame Compatible. The left and the right eye images are down-sampled vertically, and placed in the top and in the bottom of the frame.

Full High Definition 3D (FHD3D) Format This format is a lossless 3D format that provides true HD.

This is a Frame Packing format because the transmission consists of a combination of frames formed by the frame of the left and right channels. These combination of frames are stacked using the Top-and Bottom-3D format and there is 45 pixels of vertical separation between both frames. In the case of 1080p progressive frame would be 1920x2205.

This is a new format defined under the HDMI 1.4 specification and it is incompatible with HDMI 1.3 devices since they were not designed to handle a single frame with such large dimensions.

5.2.3.2 HDMI 1.3 capable 3D formats

Through HDMI 1.3, 3D content can also be transmitted using the Side-by-Side Half as an example.

The frame sequential 3D Format It is how all 3D devices work in representing images and it is also the way in which the projector used shows and needs receiving the video stream. First of all, one frame belonging to an eye is transmitted. Secondly, the next frame related to the other eye is transferred.

5.2.4 Devices

Most devices that can be found on the market work with two interfaces that refer to the left and right channels. The signals provided by these interfaces are combined into a format that complies with the HDMI 1.4 specification or, in a format supported by HDMI 1.3.

An example of this device is the Matrox MC-100. This device is able to combine and synchronize the input frame in order to form a frame format side-by-side, over/under, or frame packing (HDMI 1.4a) at a rate of 60Hz. As it is said before, all these devices are thought to be used alongside an end device. These devices are capable of managing many formats as well as displaying the content in a frame sequential way.

Moreover, there are other devices on the market with the option of converting 3D formats offered by the previous devices in a frame sequential format. An example is the Optoma 3D-XL. The limitations listed in current devices have its objection into the resolution, which is reduced from 1080p to 720p to get a frame rate of 120Hz.

Using these two devices, it would be able to record and project a flow of 3D images without any problem if the delay was not the great concern.

5.2.5 3D in real time

In the system developed, the 3D recording and processing method was subordinated to the projection method.

It was based on an image capture through two sources (2 cameras) that provided images for the left channel (left eye) and the right channel (right eye) to form a stereoscopic image.

Specifically, a pair of Toshiba IK-HR2D cameras were placed close and aligned in the same plane. This type of camera is characterized by its small size and a high performance (Appendix A). One of the main features is that this camera can capture 1080p video in a progressive way at a frame rate of 60fps.

The capture of images should not interfere with the user experience so that the cameras had to go unnoticed without compromising the quality of the system.

The combination of these sources is the aim to provide 3D signal which was screened using the F35 projector AS3D of ProjectDesign (Appendix A). This projector is capable of displaying stereoscopic images sequentially at 120Hz. Therefore, it can manage only one sort of 3D format, the Frame sequential format.

Problems to overcome

Formation of the stereoscopic frame As it has been mentioned before, the image capture is given by the use of two cameras. The formation of the 3D sequence (L-frame, R-frame) must be synchronized on time.

Moments of synchronization Normally, when trying to record a video in 3D, the cameras should start recording at the same instant of time. In other words,

they must record the same images. For this reason, many cameras for recording 3D have a synchronization signal (genlock) between them, or these cameras are post-synchronized when editing the film.

This was not the case of the cameras used for the test which are intended to 2D world and real-time recording.

However, the synchronization was provided at the time of forming the new frame by an external device.

Merging channels The external device that was used to blend the inputs and to provide this frame synchronization was the EZblend121 by Westar Display Technologies. It is a multipurpose device capable of other video treatments.

The limitation of this product is the inability to work with 1080p input signals to provide this stereo mix output. Nevertheless, it was used with 720p input to finally obtain the frame sequential format at 120Hz.

Considerations The insertion of devices to get the 3D effect introduced delay. These devices are intended to handle information transmitted in real time. Therefore, due to the constraints of our application, this solution is not viable. A hardware implementation may be made, but the idea was rejected just because this is a high-consumption process, and it was also out of scope of this thesis.

It was noticed that there was a minimum distance to perceive the 3D effect. If the objects that want to be recorded are placed at less distance than the required length, the 3D effect is not perceived. Instead of this, two images slightly separated appeared.

Regarding the frame synchronization method, no strange artefact appeared, for instance, a blur.

Chapter 6

Musical Rehearsal

A telepresence system is a multipurpose communication tool that can be used in many situations and activities. In this chapter, several tests and subjective perceptions are presented. They were gathered during the preparation sessions. This methodology was made to calibrate the constructed system and evaluate their possibilities as a tool for teaching music. In fact, the focus stem from musical rehearsal. Therefore, participants in pilot tests were a conductor, and they had several people interacting as a choir.

6.1 The system

The system was built between two rooms (picture 6.1). Thus, four links were set up in order to provide the audio and the video channels, two for the audio and two for the video. Respecting the audio links, analog signals were used as a mean to achieve almost zero delay between the two rooms. In the case of the video, the camera and projector of a link were connected through an optical link. It was necessary to rely on optical transmitters, which were made up of laser, and a receiver to perform that fact based on a photo-converter. In the other link, an unshielded twisted pair (UTP) cable was used through an DVI-D adapter.

The measured delay was according to 3.2.3.1. Consequently, an imperceptible increasing of the delay was noticed because of the electrical-optical converter and the electrical adapter of each link. According to the measurements, this increment was less than 1ms (the method used cannot compete with an accuracy below this value).

The conductor's room (picture 6.2) had as an aim to provide a near-natural sound. Therefore, the audio link, which was constituted of several channels, was connected to an analog audio mixer to overcome this purpose and to not introduce more delay. The mixer could introduce digital effects to try to change the acoustics of the room. Furthermore, two loudspeakers were put at both sides of the screen to provide stereo sound.

In the other end, the main aim was sound recording. In favor of achieving this

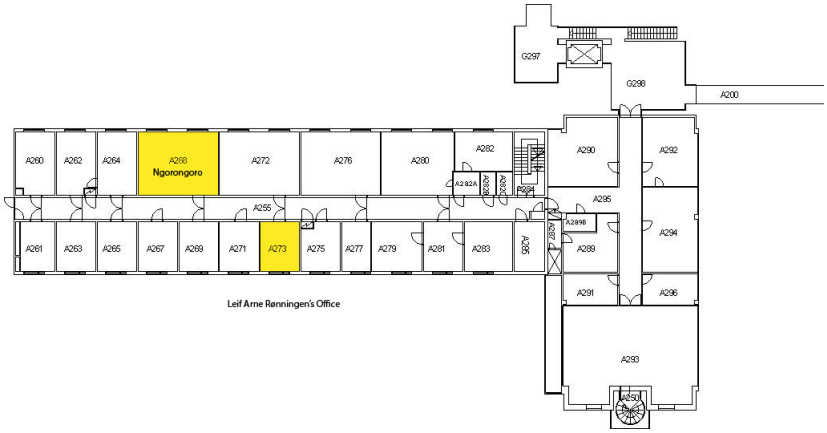


Figure 6.1: Rooms connected via Telepresence links

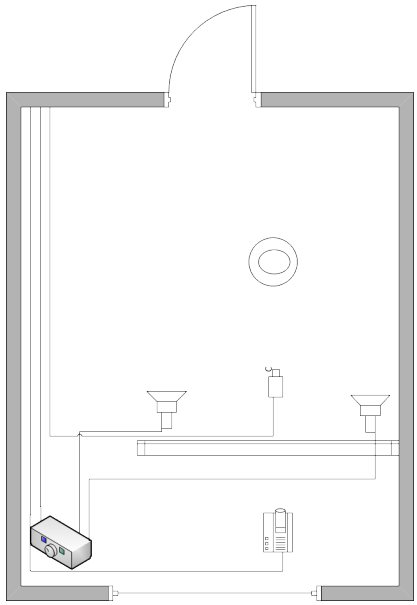


Figure 6.2: Conductor's room

fact, a pair of condenser microphones were placed in front of the musicians and at both sides of the screen.

Respecting the video, the cameras were placed between the screen and the person or people who were being recorded using a tripod. In order to not block the projection, a background projection was done, screening the content through a translucent screen and flipping the output horizontally.

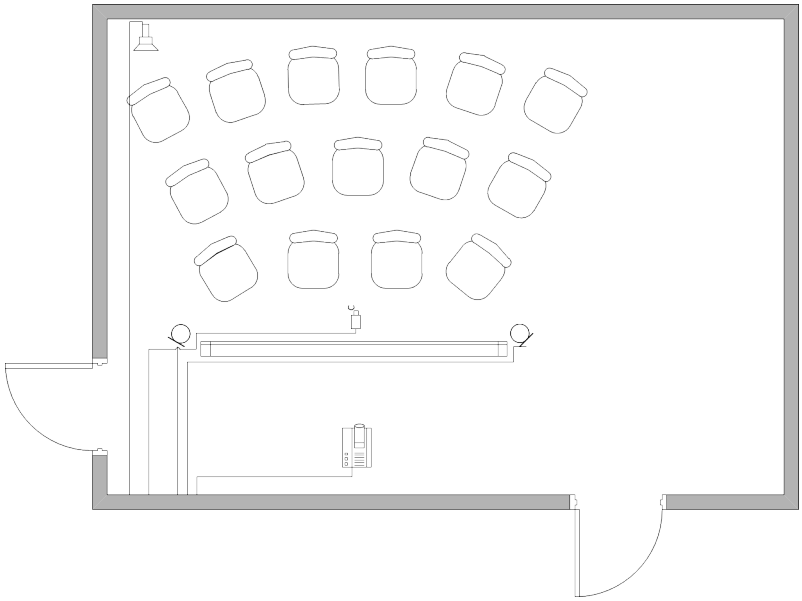


Figure 6.3: Musician's room

6.1.1 Problems

The used cameras are characterized by a small size. This size is according to the sensor-chip size. The problem of this type of sensor is the difficulty on light capture in low light conditions. Therefore, the electrical noise is present in the recording, being a bit annoying. The aperture and shutter time were optimized to capture the maximum amount of light to optimize the recording. However, it was not enough.

A way to counter this effect consists in using artificial light, but not just any light pattern. Normally, in professional environments DC powered light is typically used. The aim of this solution lies in avoiding recording the flashing lights produced by alternating current, since this produces a sweeping effect on the video. Another cheaper option is to use AC powered light but with an emitting pattern that produces a constant light, or another pattern in which fluctuations are not recorded.

Besides this, from a technical point of view, some drawbacks were possible to detect. For instance, the separation of the audio and the video signals on different channels, and the reproduction of these flows in separate devices produced an incoordination between video and sound. In fact, the delay introduced into the video channel is due to the camera processing time, the electrical-optical-converter devices and finally, the processing time onto the projector.

In the other link, the delay is imperceptible owing to the fact that the audio signal is transmitted analogically.

Nevertheless, this minimum discoordination was unnoticed by system users. In fact, according to Steinmetz (1996), a variation of -80ms and +80ms does not

produce a lip synchronization error.

6.2 Introducing delay

A way to study how the delay affects the personal perception consists in introducing it artificially.

The introduction of delay into the signal would be based on the principle of storing samples in a buffer for a while and forwarding them again. The easier way to achieve this effect in the case of audio is to digitize it. The audio signal would be sampled at a proper frequency and buffered. Regarding the video signal, it would be used the same principle, but in this case, the signal is already digitized.

This sort of data processing must be done by a standalone device, that is to say, without the execution of an operating system. A proposed solution can be seen in Appendix D. In case of using a computer with a data acquisition card, the communication between the driver and the superior layers must be broken. If the operating system is running and sharing the resources among other threads, the delay cannot be constant.

There are many digital audio mixers that can control the amount of delay. These delays want to be added to the signal in gradual steps. Another possibility is the independent devices that are being used to synchronize the video and audio when broadcasting.

It is also possible to use a pair of encoder/decoder as an alternative method. It was mentioned in 3.2.3.6 that the decoder device used is able to buffer the incoming packets. These packets content the video and audio data. Hence, changing the size of this buffer modifies the delay.

The only difference is that a combiner is needed to merge the audio (digital) and video signal into one signal.

6.3 Scenario analysis

In a musical rehearsal, the communication between the musicians and the choir is made as a normal communication situation through vocal communication and gesture perception. However, these communication channels do not have the same importance from a musical point of view.

Regarding the conductor, the aim is to guarantee that musicians' interpretation is being done according to his/her viewpoint and guidelines. Hence, a proper reception of sound is needed to make a comparison between what is received and what is expected. The musicians also need to watch the conduction to guide their interpretations.

The other channels, which are drawn as a dashed line in figure 6.4, do not require such attention, that is to say, there is more tolerance.

In fact, the delay experienced by individuals is different depending on which role are performing.

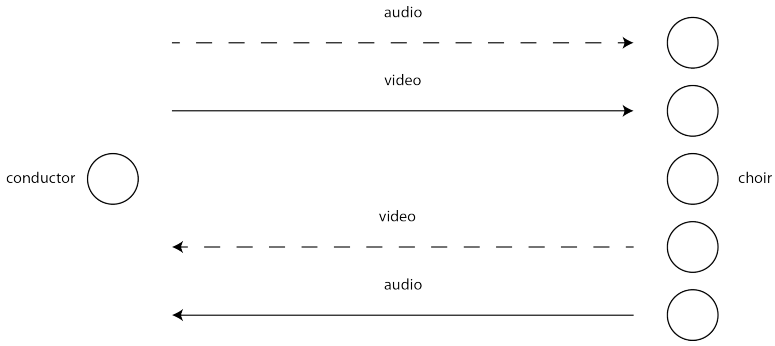


Figure 6.4: Musical rehearsal important links

Concerning musicians, the delay appears just when there is an interaction with the conductor regarding both ends. Otherwise, musicians are interpreting respecting what it is presented in the screen, for instance, when watching a video. However, the delay suffered by the conductor is longer. The conductor needs to receive the sound emitted by musicians, whereby the delay that the conductor will experience will be the video connexion from the conductor to musicians plus, the sound link from musicians to the conductor.

6.4 The system with variation on delay

Once the starting configuration was running properly, it was decided to increase the possibilities of testing. The decision pretended to be able to control the delay between ends for both audio and video.

In the audio case, the increase of delay was made using two different solutions regarding which end is transmitting data. As it is mentioned in the previous section, all links do not have the same importance according to the paradigm which was being tested. Regarding the audio, the main important link goes from the choir to the conductor. Because of this, in the configuration scenario presented in figure 6.2, the audio-end devices are of a high-quality. Furthermore, in order to increase the delay a digital mixer capable to achieve this fact from 0 to 300ms was used. By using this digital mixer, there was not any problem with the audio interfaces.

Respecting the other audio link (conductor-choir), the delay was tried to be added using a computer running Linux. That was done using an audio-capture card and a small tool ran via script (Appendix E). However, the control over the delay was almost none. As it was explained above, the operating system must not take part of this process because it would run other processes, assigning them priorities to share the CPU use.

In the case of video, owing to the fact that no available device on the market was found to perform this task, it was decided to use the delay introduced by equipment in order to have a completely control. Nonetheless, a pair of computer with HDMI capture cards running another tool into a Linux system (see Appendix E) were

used to evaluate their performances.

6.4.1 Configuration measurements

To be aware of real delay and performances of these solutions, several measurements were taken.

6.4.1.1 Audio

The delay introduced by the digital mixer was extremely accurate. In fact, the only task that was done was to measure the delay introduced by the analog-to-digital converters and vice versa to take it as a minimum delay. An oscilloscope to measure this magnitude was used. Furthermore, this value could be incremented using the delay option provided by this device.

The used configuration is sketched in 6.5.

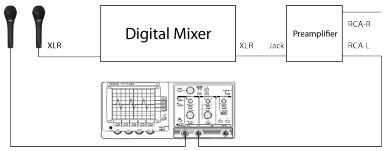


Figure 6.5: Digital mixer delay measurement

The value obtained without setting the delayer capability was 2.4 ms.

Regarding the other audio link, the performance was measured using the configuration depicted in figure 6.6. But as it is mentioned, this tool should not be used to introduce delay. Actually, the delay had no relation even when the set delay values were about 1 second. The value remains constant once the application is running but the set value is not related with the real delay obtained. Hence, in order to use this tool, the procedure was to set a value, to run the applications and to perform the measurements to figure out which delay was obtained. This task was repeated until getting the desired value, tweaking the input parameters of the script.

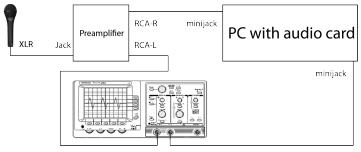


Figure 6.6: Rooms connected via Telepresence links

6.4.1.2 Video

The configurations used were taken from 3.2.3.1 (initial configuration), 3.2.3.2, and from a new case. This new scenario consisted in using the same camera than in

3.2.3.3, a high-end camera but shooting 720@60fps video. The delay introduced was 70ms. Consequently, the delay was known and constant to these cases.

The last case was based on using capture cards to disclose the new delay behaviour and to evaluate their performances. The measurements were done running an Ubuntu Linux computer, using a low-latency kernel. The values obtained had a big variance as it was expected. In order to try to diminish this fact, the priority and scheduling discipline of the script were modified. However, these measures did not improve the performance, obtaining the same variance. One explanation to this fact stems from the video output of the used tool that was using the OS's desktop environment. This variance appears because there are many running processes to show the output.

6.5 Pilot tests

The purpose of these sessions was to test the system in order to know which were the possibilities that the system could provide. Moreover, the drawbacks and weaknesses were analysed to figure out a solution before the general rehearsal.

During the first two sessions, three types of exercises were practiced to interact with the system and to evaluate the sensations perceived about, being a tool to compare between several scenarios.

Conduct and clap It consists in a clapping test which was conducted by the conductor whereas the other participants were trying to keep the rhythm. Just one clap per pulse. The tempo was varying according to the conductor gestures.

Conduct and sing In this test, the participants sang the "Frère Jacques" song using the syllable /ta/ and the conductor was speeding up and slowing down the tempo on the fly.

Clap and clap This time, the conduction was carried out by clapping. Furthermore, the first phase was made clapping vertically whereas the next phase was performed horizontally.

6.5.1 First session

The participants in this session were the conductor, a member of the choir and two people. The configuration of the system is described in 6.1. Each individual was placed in its right place in order to start with the session. The conductor is in an independent room and the other people were together. The first approximation was performed to be in contact with the system.

Therefore, the first and third clapping tests presented above were practiced over the system. That configuration was zero delay for the sound and 27ms for the video, which was the delay introduced by this equipment configuration.

Consequently, all members discussed the first impressions and their feelings. The same procedure was repeated but being the participants with the conductor in the same room in order to appreciate the difference. Afterwards, the previous tests were performed, using the system and adding the second test (Conduct and sing).

As a first impression, all participants agreed that feelings were quite similar between both situations (same room and over the system). However, the conductor said that at the first time, she was conducting a bit unsure, having the need to speak more clearly. A possible explanation lies in that she did not change her mentality of observing the people through a screen. The second round, she was more confident with the system and she conducted in a normal way.

From musician's point of view, it was appreciated that when the tempo was going very fast, the video was not been recorded properly. It seemed to be stuck because the frame rate (60fps) was not enough. Furthermore, the screen frame helped in the conducting task due to the fact that the video was not showing the entire body of the conductor. In fact, the bottom line of the frame could be used as a reference point. Besides this, all participants agreed that having a camera in front of the screen was a bit annoying.

6.5.2 Second session

Several new scenarios were prepared with the intention to analyse how the quantity of delay affected the system. That fact was done following the methodology explained in 6.4. Before doing a subjective evaluation of each scenario, all available possibilities were tried later on to decide which one was worthy of analysis, or which configuration was introducing something new regarding the previous configuration.

The configurations that were analysed are presented below:

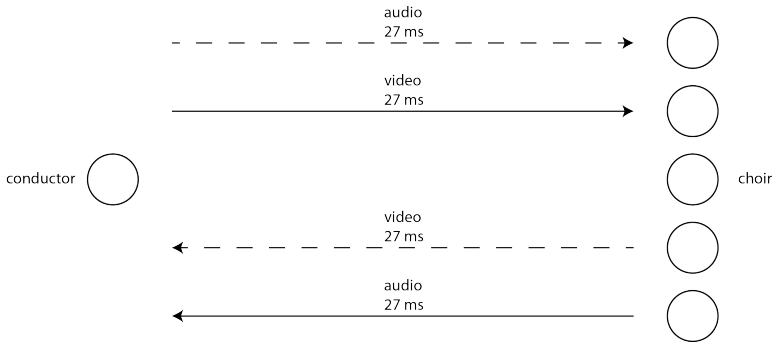


Figure 6.7: Scenario A

The session started doing the three mentioned tests in the same room in order to take as a reference this configuration. The tests were repeated in each scenario later on. The participants were the same than in the previous session.

From musician's point of view, no change in the configuration affected their

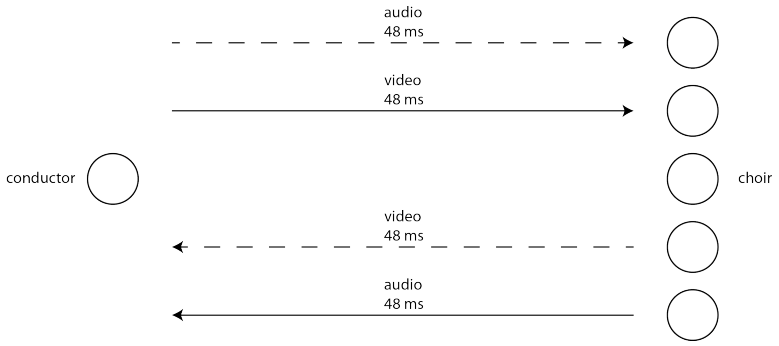


Figure 6.8: Scenario B

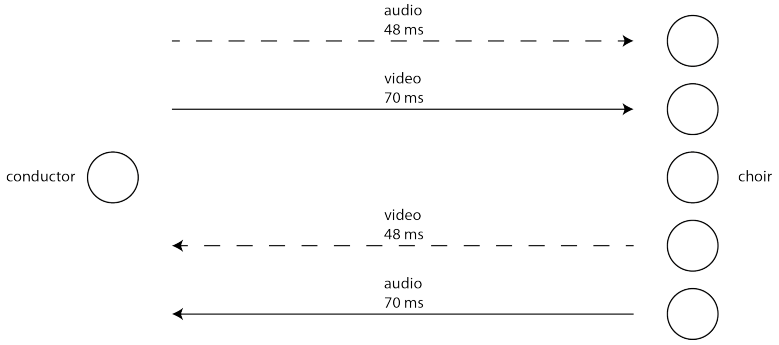


Figure 6.9: Scenario C

perception because they reacted after watching the screened images. In principle, delay does not matter in this case. Instead of that, the conductor's perception was altered due to fact that the delay appeared in both directions (see 6.3).

In Scenario A (figure 6.7), the delay was not noticeable. Nevertheless, in Scenario B (figure 6.8) no problems were presented while conducting, but the delay was perceptible in the clapping tests. In fact, when tempo was speeding up, musicians were running late and when the tempo was slowing down, they were in sync. In Scenario C (figure 6.9), the delay was manageable similar to what is perceived when conducting a large choir. In the 'clap and clap' test, the perceived sound was out of sync in conductor's end. In addition, the sound seemed to have mordents. In table 6.1 a summary of the test results in the different scenarios is presented.

6.5.2.1 Considerations

The problems that arose in the first session like the visual blocking of the screen, were no longer an issue. However, in Scenario C a bigger camera was used which reduced the contact with the person that was blocked by this camera. No reduction was noticed regarding other people.

	Delay perceived by the conductor	Conduct and clap	Conduct and sing	Clap and clap
Scenario A	54 ms	no noticeable	no noticeable	no noticeable
Scenario B	96 ms	no noticeable	no noticeable	perceptible
Scenario C	140 ms	manageable	manageable	out of rithm

Table 6.1: Conductor's feelings into several scenarios conducting mentioned tests

Another reported effect was the increased feeling of remoteness when the person that appeared in the video did it with a smaller size.

The video was also shoot at 30fps, reporting smudged movements by the participants who did not notice this fact. It is impossible for the conductor to compensate this fact. Hence, shooting at 30fps was not enough at least for this sort of interaction.

Despite these drawbacks, a fast learning curve was experienced by all participants. Furthermore, the system became more human-friendly when time was passing along.

6.5.3 Third session

During the last session, participants were singers of a choir. On the one hand, the room destined to the tests was not ready for musical performances, being too small and with very poor acoustics. Therefore, that fact can influence the judgement capabilities. On the other hand, the session was splitted into two parts. The first steps consisted in repeating again the previous three interactive types of exercises over several scenarios. In order to speed up the session and gather straightaway information to avoid the loss of data, participants had to fill out an online survey after each scenario performance.

The second part was devoted to discuss the screen size and the visualization of several videos. In these videos, several conductors were conducting the same piece of music (Grieg's Ave Maris Stella) while the components of the choir performed according to those videos (see picture 6.10). Moreover, they filled out a survey after viewing each video as well.

In Appendix F, the results and the questions that appeared in the surveys can be found as well as comments.

6.5.3.1 First part

The scenarios that were tested are presented in table 6.2, following the same methodology applied in previous pilot tests.

It is noteworthy that group members have never sung together before. In the case of the conductor, she took notes on her findings.

Figure 6.10: Conducting Grieg's Ave Maris Stella trough a video recorded previously

Using the acquired information, several observations are presented below even though there is a bit of disparity in the answers.

From the conductor's point of view, the feelings were alike to previous tests. As in former sessions, a local rehearsal was done before starting with the other scenarios to establish a reference (see picture 6.11). The conductor already noticed a less precision in rhythmic tests in Scenario 1 becoming this feeling further aggravated in the next cases. This difference, in contrast to previous sessions, can rely on the number of cooperators who were participating at that time. In Scenario 3, no problems arisen in the conducting exercises whereas in the clapping tests it was difficult to be in sync. Furthermore, sound and audio were in sync, without reporting any issue by the conductor.

The first time that she perceived a disturbing feeling while conducting the clapping tests was in Scenario 4. Moreover, a difference between the clapping techniques used by participants increased the difficulty to analyse the performance. Apart from that, no problems conducting "Frère Jacques" song (conducting and singing) were noticed. Both multimedia channels were still in sync. Finally, in Scenario 5 the delay was evident, being possible to work with steady tempo. Nonetheless,

		Video Quality	Video delay	Audio delay
Local Rehearsal 1				0
Scenario 1		1080p@60fps	27ms	27ms
Scenario 2		1080i@60fps	48ms	48ms
Scenario 3	(conductor-choir)	720p@60fps	70ms	48ms
	(choir-conductor)	1080i@60fps	48ms	
Scenario 4	(conductor-choir)	1080i@60fps	48ms	48ms
	(choir-conductor)	720p@60fps	*variance*	
Scenario 5	(conductor-choir)	1080i@60fps	48ms	48ms
	(choir-conductor)	720p@60fps	*variance*+50	

variance= *54ms (minimum value), 120ms (maximum value)*

Table 6.2: Scenario configurations of session 3

Figure 6.11: Conducting a local rehearsal

speeding up the tempo and slowing it down was very complicated. In fact, the conductor had to be advanced rhythmically, focusing all her efforts in what would

come instead of analysing the feedback received. Though, it was still possible to conduct songs.

Regarding the singer's point of view and according to the survey results (see Appendix F), a degradation in musical aspects is manifested.

In relation to question Q1i (see results in F.1), in Scenario 2 and 3 many people responded 'Neither I Realized'. This difference respect to other scenarios lies in the camera location. It was placed in a different position. It was no longer in the centre of the screen, instead of this, the camera was situated below the screen view. Nevertheless, a criterion discrepancy (subjective data) was observed regarding Scenario 4 and 5. In both scenarios the same configuration was used.

Concerning the quality of image, the participants were aware of an enhancement in the image quality in Scenario 3. That difference corresponds to a high-end camera that was used in the conductor-choir link to set the right delay and contributing with this change.

With respect to phrasing/musical expression (question Q8i (F.8)) and quality of dynamics (question Q7i (F.7)), a degradation compared to other scenarios is produced. This fact seems to be related to musical problems that the conductor is bearing. That is to say, the conductor is the person who really experiences the disadvantages of delay. The choir only reacts to a video which is influenced by the feedback received from the performances that can be seen on it. Therefore, those problems can be perceived by the choir.

Questions Q5i(F.5) and Q6i(F.6) have not had a big influence, which means that they were not such a good questions.

In general, a reduction in performances is appreciated in all scenarios.

6.5.3.2 Second part

One more time, a local rehearsal was performed in order to have a reference before evaluating the other scenarios. These configurations are shown in table 6.3.

After doing a break, the evaluation of the different scenarios started again. The Scenario 1 was repeated, calling it now Scenario 8. As a surprise, the responses regarding these configurations were more favorable than in part 1. Therefore, an improvement in almost every aspect was appreciated. This fact can be related to the learning curve of the system. The next three videos were recorded by the same conductor. A great difference regarding the perceived communication (F.12 and F.13) is shown between Scenario 10 and 11. That discrepancy can be attributed to the detail that in Scenario 11 the conductor was looking directly at the camera. This circumstance also takes part in Scenario 9 but the fact of being performed after Scenario 8 reduced its repercussion.

Another interesting point involves the video quality of Scenario 14. As it can be appreciated in F.11, the fact of using an interlaced acquisition mode is not reflected. Actually, just two people reported blur artifacts with fast movements. Anyway, this fact plus a less precise conduction is reflected in the ability to find the right tempo (see results in F.18).

	Video Quality	Comments
Local Rehearsal 2		
Scenario 8	1080p@60fps	Real time
Scenario 9	1080p@24fps	Conductor looking at the camera
Scenario 10	1080p@24fps	Conductor **not** looking at the camera
Scenario 11	1080p@24fps	Conductor looking at the camera
Scenario 12	1080p@29.97fps	Conductor 2
Scenario 13	1080i*@25fps	Conductor 3
Scenario 14	1080i@24fps	Conductor 3, conduction less precise

* *the video was desinterlaced using a video editting tool*

Table 6.3: Scenario configurations of session 3

6.5.4 Screen size

During these pilot tests, the possibility of using recorded video as a tool for practice was also tested. As it is mentioned, people who are in musicians' side have to react to the movements of the video.

In principle, common people do not have at their homes a big screen to show a person recorded in a natural size. Furthermore, several configurations were tried to obtain their impressions.

After interviewing musicians, it was clear that the conductor should have filled as much screen as possible. However, when the screened person was larger than her natural size, which was the optimal size, a strange feeling persuaded the cooperators as they were seeing a person bigger than her real size. Furthermore, the fact that the conductor was far away from the camera produced more discomfort. Regarding smaller screen sizes, the singers pointed that they resembled the same common distance as if they were in a cathedral.

6.5.5 Treating the system

In this type of scenarios, having some considerations about the recording is very important.

In a common rehearsal, the conductor is paying attention to the musicians, looking at their eyes directly. The person who is being recorded must treat the camera as a person if the same sensation wants to be transmitted. Clarifying this, the conductor must look at the camera as he/she would look at a musician. Instead of this, a sensation of awayness would be transmitted to viewers. Besides this, this fact cannot be maintained for a long period of time, due to the fact that viewers can feel a bit observed. In real life, the attention to one person is focused just for a short period of time.

Chapter 7

Discussion

This chapter contains a discussion in each section of the obtained results. Since the work has treated several portions, this discussion is also divided into several topics, making it a completely independent analysis and contributing with some ideas for future works.

7.1 Discussion of the delay study

This section is related to the content and the work presented in Section 3.2. Therefore, two different areas can be distinguished.

Proposed delay-measuring method

The proposed method is intended as a low cost and simple methodology to estimate the optical delay. Because of this simplicity, it presents some drawbacks that can be solved as it is explained below. Besides this, the person who is taking these measurements needs to bear in mind some considerations.

The measured value depend on which colour is used as a stimulus signal and how the colour wheel composes that colour (see 3.2.2.4). Moreover, the asynchronous working way of the method thwarts the accuracy. Thus, the way it works consists in taking the measurements manually, and using the markers of the oscilloscope to obtain the displayed values on time. In order to increase the accuracy is crucial to take that values for a long period of time, increasing the chance of obtaining the limit values.

Future work

According to the proposal 3.2.6, an automation of measurement process would be gotten, leaving aside the fact of taking the measures manually. Besides this, an increasing in the accuracy would be manifested, since it would be possible to leave the equipment working without human interaction for an enormous period of time.

Another possibility to avoid the presented drawback would be to create a device capable of detecting when a frame is sent out and when it is received. This tool should work similarly to how a network traffic analyzer works. Instead of focusing in packets, frames would be the target.

Measurements of a real telepresence system

The different local scenarios presented and the measurements in real networks disclosed that the greatest amount of delay in current systems is added by the equipment used instead of the network. The recording and the projection equipment were the main responsible tools.

Hence, posterior works would be focused on reducing this issue.

However, the values obtained are much better than either current commercial telepresence system when the buffer size is small. Accordingly, this device configuration (Toshiba camera IK-HR2D, Hitachi HU-200EI, Hitachi HU-200DI and ProjectionDesign F35 AS3D projector) can be used to develop a proper telepresence system intended for applications demanding low latency. The first drawback is the high cost. This equipment is not cheap and it limits the scope of use. Another important problem is the configuration. It does not exist a user-friendly application to configure the communication. It has to be done manually, changing the network parameters of the encoder and the decoder. However, this is a good point to start.

New network architectures are being proposed like Distributed Media Play (DMP) as a way to rebuild completely the protocols used in transmissions as well as the role that the router performs in the network (Arne, 2007). The idea consists in giving more intelligence to the network. Nonetheless, as it can be noticed, the end-devices used to capture and project the scenes need a drastically reduction on delay.

7.2 Reaction time and frame rate conclusions

In Chapter 4 several tests were done to figure out which was the relation between the frame rate of a video and the time response to the events presented in that video.

After analysing the results, it was concluded that the mean reaction time to the test 4.3.3 is about 295ms. In the published works, it is said that the mean RT to detect a visual stimulus is approximately 190ms. It is worth to mention that this value refers to tests in which the individual under test just had to press a button instead of emitting a sound. Hence, the fact of using the vocal apparatus to produce that sound provokes that difference. In addition, a key concept that reinforces this theory relies on the fact that the same test (only a few samples) was performed by treating the microphone like a button. The tester had to strike a blow with his finger to the microphone, imitating the act of pressing a button, instead of making a sound. As a result, an improvement in the measurement was automatically observed, close to those button-pressed tests.

Regarding the frame rate and according to the results obtained, an improvement in the RT was detected when the event occurred, having the tester enough time to predict its trajectory. That is to say, the more higher decision time, which provides more information, the easier it is to know when the event will take place. However, when the event occurs just for a short period of time and the predictability is low, this effect disappears.

7.3 Discussion about depth perception

This discussion deals with the statements presented in Chapter 5.

The work has attended to increase the depth perception in telepresence systems. The developed system was based on using active stereo 3D via shutter glasses and a frame sequential projection. After experimenting with this solution, an increase of depth perception became manifested. However, this does not solve the pointing problem that was exposed at the beginning of that chapter.

The fact of using just one point of view (two cameras to form one 3D signal) is not enough to overcome that situation.

This is a good solution when just one person is intended to experiment the 3D signal, being the only receiver. For instance, it would be a good solution for a conductor which is receiving the signal from a choir. Additionally, this does not apply when several people are watching the same output. This is the same effect that it is experienced when watching a film. Everybody is watching the same images, same gestures.

Taking the tests of the choir as an example, when the conductor is pointing to the sopranos, each component of the choir would be watching the same images though with a depth perception.

Future work

A purpose that must be tested as a future work would be using a multiview system. In this sort of systems, different points of view are shown depending on where the observer is placed. However, this type of systems are not still enough developed having viewing problems. Nevertheless, the applicability must be tested.

7.4 Discussion about pilot tests

Several tests were conducted to evaluate the viability of the developed system in the field of musical rehearsals. The content related to these tests and all considerations are described in Chapter 6.

The pilot tests were performed using several considerations whose peculiarity was the introduction of different delays into the video and audio links.

After those sessions, it was noticed that the conductor was able to even tolerate a delay about 140ms (round-trip delay). On the one hand, these tolerable effects also appear in (Chew et al., 2005), where two pianist players tolerated 50ms (one-way). Nevertheless, that fact depends on conductor's skills, experience and it also

depends on the type of music. As it is commented in this chapter, the main problems arose when rhythmical tests were carried out, and also when the tempo was fast. This fact agrees with the researchers conducted by Chafe et al. (2004). Nonetheless, practice increases tolerance.

On the other hand, the conductor makes efforts which influence its conduction in order to bear the delay. The effort is perceived by participants. This is the channel whereby participants receive the feedback. Thus, the performance will be diminished.

A technique to increase the closeness between the conductor and the musicians consists in looking directly at the camera. Therefore, this direct view transmits the feeling that the conductor is looking at the eyes of the individual who is watching the video streaming. It is also applied when recording a video.

Participants reported that their feelings were not the same sensations that they had in a local environment. In any case, after some practise, the perception improved. Clarifying this, the first impression is always reluctant. However, when players start to get used to the system, their outlooks improve.

Another important aspect to enhance the experience is based on illuminating the room with a light. The light does not emit a variable pattern in order to avoid producing a flickering in the video. The effect produced is annoying, increasing the tiredness of players.

Regarding the optimal screen size, it was clear that it corresponds to the one that provides the size of the displayed person as in real life.

Concerning the frame rate, recording at 30fps is a low frame rate for this sort of activity. Most of the people appreciated the annoying artifacts that appeared when rapid movements were recorded compared to the same video movements at a higher frame rate. According to the tests, a frame rate of 60fps seems to be enough.

Future work

New tests should be carried out, taking a completely control over the delay instead of taking advantages of the delay introduced by the equipment. In this way, a more accurate threshold can be obtained. Furthermore, a qualitative data was gathered after analysing the surveys. New enhancements can be added to next surveys based on asked questions. In addition, new focus points have come to light. Besides this, new situations like musical collaborative spaces, in which musicians are placed in several locations, could be tested keeping in mind that tolerance values will become more restrictive.

A way to compare recorded videos with a real time rehearsal consists in shutting the video and audio channels down. These audio and video channels would not be received by the conductor in order to eliminate the feedback. Thus, this situation is equivalent when recording a video. The only difference is that the video will be watched at the same time that it is created. Furthermore, participants do not need to be aware of this phenomenon.

Chapter 8

Conclusion

Several aspects related to the perceived quality of a telepresence system have been presented in this thesis. In fact, a method to measure the optical delay has been proposed and tested in different scenarios, either using public networks. The response time has been another aspect treated. The mean value (295ms) was obtained regarding visual and vocal response time according to the measurements done. Moreover, the fact of using a high frame rate improved the response time when participants could predict when the event would occur. In order to carry out the study of this point, several people participated selflessly in the tests. Moreover, a search in the current marked was done to be able to show 3D video in real time and to tackle the problem of perception in this type of system. The proposed implementation did not enhance the perception enough.

The collaborative help of people was also an important condition which was necessary during the pilot tests that were performed over the built telepresence system. A threshold in conduction perception was obtained as well as personal evaluations regarding the system. Furthermore, a demonstration was done to a group of people who were not involved with this project. In this demonstration all drawbacks and difficulties that arose in the project were explained as well as which were the possibilities and the proposit of this research.

Finally, a discussion presenting the achievements and results was done. Therefore, several proposes to improve the results obtained and new suggestions for future work were also mentioned. In fact, the pilot tests will be continued under the supervision of the co-supervisor of this thesis (Otto J Wittner) and the conductor who participated, Vivianne Johnsen Sydnes.

Bibliography

L. Arne. The dmp system and physical architecture, December 2007. URL http://www.item.ntnu.no/~leifarne/The%20DMP%2014Sep07/The%20DMP%20System%20and%20Physical%20Architecture.htm. Web page retrieved 16.05.2012.

A. Barbosa. Computer-supported cooperative work for music applications. Master's thesis, 2002. URL files/publications/0fbbc9-PhD-Abarbosa-2006.pdf. Web page retrieved 20.04.2012.

John Bartlett. Telepresence: Beautiful and Expensive. *Business Communications Review*, 37(6):20–25, June 2007.

C. Chafe, M. Gurevich, G. Leslie, and S Tyan. Effect of Time Delay on Ensemble Accuracy. In *Proceedings of the International Symposium on Musical Acoustics (ISMA2004)*, Nara, Japan, March–April 2004.

E. Chew, A.A. Sawchuk, R. Zimmermann, V. Stoyanova, I. Tosheff, C. Kyriakakis, C. Papadopoulos, A.R.J. François, and A. Volk. Distributed Immersive Performance. In *Proceedings of the 2004 Annual NASM Meeting*, San Diego, CA, November 22 2004. URL http://nasm.arts-accredit.org.

E. Chew, R. Zimmermann, A.A Sawchuk, C. Papadopoulos, C. Kyriakakis, C. Tanoue, D. Desai, M. Pawar, R. Sinha, and Meyer. W. A Second Report on the User Experiments in the Distributed Immersive Performance Project. In *Proceedings of the Fifth Open Workshop of MUSICNETWORK: Integration of Music in Multimedia Applications*, Vienna, Austria, July 4-5 2005. URL http://www.interactivemusicnetwork.org.

S. Coren, C. Porac, and L.M. Ward. *Sensation and perception*. Academic Press, 1984. ISBN 9780121885557.

Digital Display Working Group. Digital visual interface dvi, April 1999. URL http://www.ddwg.org/lib/dvi_10.pdf. Web page retrieved 14.05.2012.

Xiaoyuan Gu, M. Dick, Z. Kurtisi, U. Noyer, and L. Wolf. Network-centric music performance: practice and experiments. *Communications Magazine, IEEE*, 43 (6):86 – 93, june 2005. ISSN 0163-6804. doi: 10.1109/MCOM.2005.1452835.

L. Handberg, A. Jonsson, and C. Knudsen. Community building through cultural exchange in mediated performance events. 2005.

Jari Kleimola. Latency Issues in Distributed Musical Performance. *Seminar on content creation*, Fall 2006.

D. Konstantas, Y. Orlarey, O. Carbonel, and S. Gibbs. The distributed musical rehearsal environment. *Multimedia, IEEE*, 6(3):54 –64, jul-sep 1999. ISSN 1070-986X. doi: 10.1109/93.790611.

Dimitri Konstantas, Yann Orlarey, Simon Gibbs, Olivier Carbonel, Jean Moulin, F Lyon, and D Sankt Augustin. *Distributed Musical Rehearsal*. Number 95. 1997. URL `http://quod.lib.umich.edu/cgi/p/pod/dod-idx?c=icmc;idno=bbp2372.1997.075`. Web page retrieved 2.02.2012.

R Kosinki. A literature review on reaction time, September 2010. URL `http://biae.clemson.edu/bpc/bp/Lab/110/reaction.htm`. Web page retrieved 16.01.2012.

Y. Lacouture and D. Cousineau. How to use matlab to fit the ex-gaussian and other probability functions to a distribution of response times. 4(1):35–45, 2008.

Y. Orlarey, O. Carbonel, D. Konstantas, and S. Gibbs. Distributed musical rehearsal : Evaluation report. July 1998.

Nathan Schuett. The Effect of Latency on Ensemble Performance. Bachelor thesis, CCRMA Department of Music, Stanford University, Stanford, CA, USA, May 2002.

R. Steinmetz. Human perception of jitter and media synchronization. *Selected Areas in Communications, IEEE Journal on*, 14(1):61 –72, jan 1996. ISSN 0733-8716. doi: 10.1109/49.481694.

R. Whelan. Effective analusis of reaction time data. 58:475 –482, 2008.

Abbreviations

ASI Asynchronous Serial Interface. 19

ATM Asynchronous Transfer Mode. 7

DIP Distributed Immersive Performance. 7

DLP Digital Light Processing. 14

DMP Distributed Media Play. 62

DVI-D Digital Visual Interface - Digital. 17, 18, 32, 47, 85

DVP Distributed Video Production. 7

EFA Ensemble Performance Threshold. 11

FEC Forward Error Correction. 20, 21, 24

HDMI High Definition Media Interface. 18–20, 33, 42–44, 51

LED Light-emitting diode. 12, 17

PS3 PlayStation 3. 33

RGB Red, Green, Blue. 14–17

RT Reaction Time. 29, 30, 35, 62, 63

SDI Serial Digital Interface. 17–20

TTL Transistor-transistor-logic. 25

UTP unshielded twisted pair. 47

Definitions

background projection Method of projecting pictures onto a translucent screen so that they are viewed from the opposite side. 33

FCCN (Fundaçao para a Computação Científica Nacional) FCCN is a private non-profit institution from Portugal which manages and operates the national NREN, Science, Technology and Society Network (RCTS). 21, 24, 25

Hitachi It is a diversified Japenese company. 17

MATLAB It is a programming environment for algorithm development, data analysis, visualization, and numerical computation. 35

Mini DisplayPort It is a miniaturized version of the digital audio-visual interface developed by Video Electronics Standards Association (VESA). 32

mini-jack It is an audio connector used to transmit analogue sound. 32

ping It is a network tool to measure the round trip time. 23, 24

process It is an instance of a computer program that is being executed. 32

RCA It is a audivisual connector. The name "RCA" refers to Radio Corporation of America, which introduced the design in 1949. 33

Rolling Shutter Refers to the acquisition method used in most camcorders which consist of sweeping line by line to record each line in different times. 15

Round-Robin Scheduling discipline in which an operating system assigns each process a fair and orderly portion of time. 32

Transition-minimized differential signaling (TMDS) It is a technology for transmitting high-speed serial data based on encoding the signal to protect it. 85

tri-level It is diferential signals. The pulse will start at the zero volts and first transitions negative. After a specified period, it transitions positive, holds for a specified period and then returns to zero or black level. This symmetry of design results in a net DC value of zero volts. 25

Uninett It is a state owned company responsible for Norway's National Research and Education Network. III, 17, 21, 24

Appendices

Appendix A

Equipment features

A.1 Encoder/Decoder

The used models are HU-200EI and HU-200DI by Hitachi Kokusai Electric Inc.

Main features

Compression method It is used the MPEG-4 AVC/H.264 method for compression and the packets are sent as MPEG-2 Transport Stream

Inputs (From the encoder point of view) HD-SDI video input

Outputs (From the encoder point of view) Ethernet and ASI interfaces

Processing time 8 ms approximately at 59.94i.

A.2 Projector

The F35 AS3D is a high performance projector.

Main features

Resolution 1080p (1920 x 1080p), WUXGA (1920 x 1200)

Stereo mode It is only available when the source video has a frame rate between 100Hz and 120Hz

Computer inputs 2 x DVI-D2 x HDMI 1.3a

A.3 Camera

A.3.1 Toshiba camera

The Toshiba camera IK-HR2D is a Ultra Compact camera (approx 1.75 inches x 1.75 inches x 3 inches) one-piece system provided with a CMOS sensor.

Output Outputs full 1920 x 1080 pixels at 60 frames per second Output selectable between 1080p, 1080i and 720p

Sensor Utilizes 1/3" CMOS 2.1 megapixel HD sensor

Output interface Digital DVI

Lens mount C-mount

A.3.2 Panasonic Camera

The Panasonic AG-HPX250 P2 (7 inches x 7-11/16 inches x 17-1/4 inches) is a Professional Handheld Camera Recorder.

Output 1080/59.94i, 1080/29.97p, 1080/29.97pN, 1080/23.98p, 1080/23.98pA, 1080/23.98pN, 720/59.94p, 720/29.97p, 720/29.97pN, 720/23.98p, 720/23.98pN, 480/59.94i, 480/29.97p, 480/23.98p, 480/23.98pA, 1080/50i, 1080/25p, 1080/25pN, 720/50p, 720/25p, 720/25pN, 576/50i, 576/25p

Sensor It is compost of a high-sensitivity, low-noise 1/3-type 2.2-megapixel U.L.T. (Ultra Luminance Technology) 3MOS image sensor

Output interface HDMI, SDI

Lens Optical image stabilizer lens, 22x motorized zoom, F1.6 – 3.2 (f=3.9 mm – 86 mm), 35 mm conversion: 28 mm – 616 mm (16:9)

Appendix B

Results of Response Time measurements

B.1 Results of Scenario B

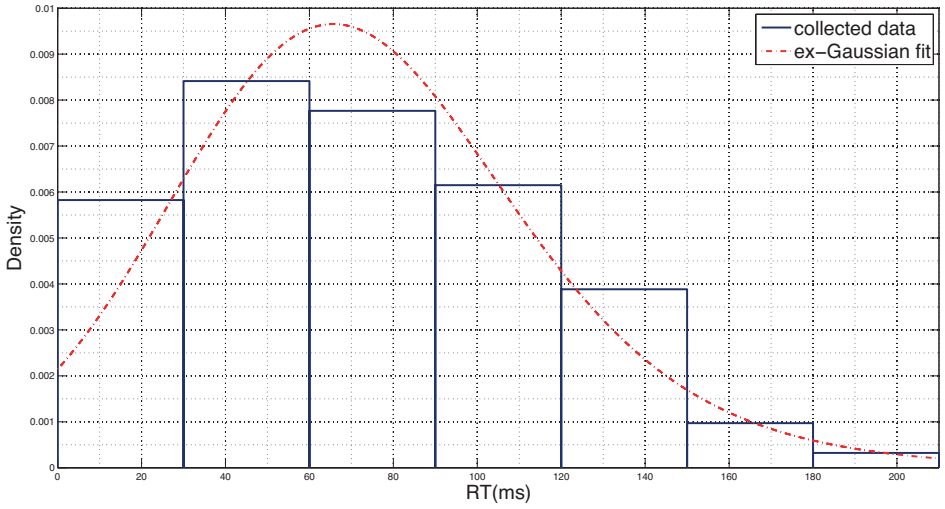


Figure B.1: Probability density function (PDF) of RT samples acquired in scenario B(4.3.3) using the 30fps recorded video

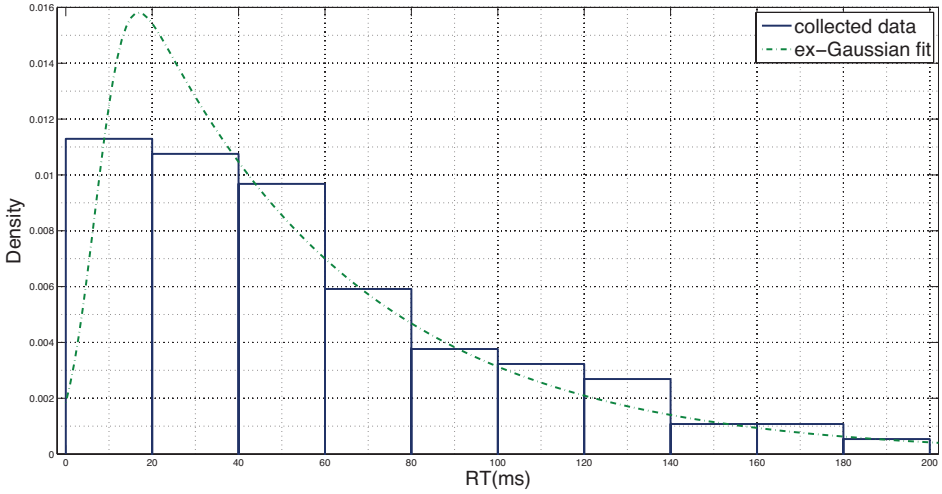


Figure B.2: Probability density function (PDF) of RT samples acquired in scenario B(4.3.3) using the 60fps recorded video

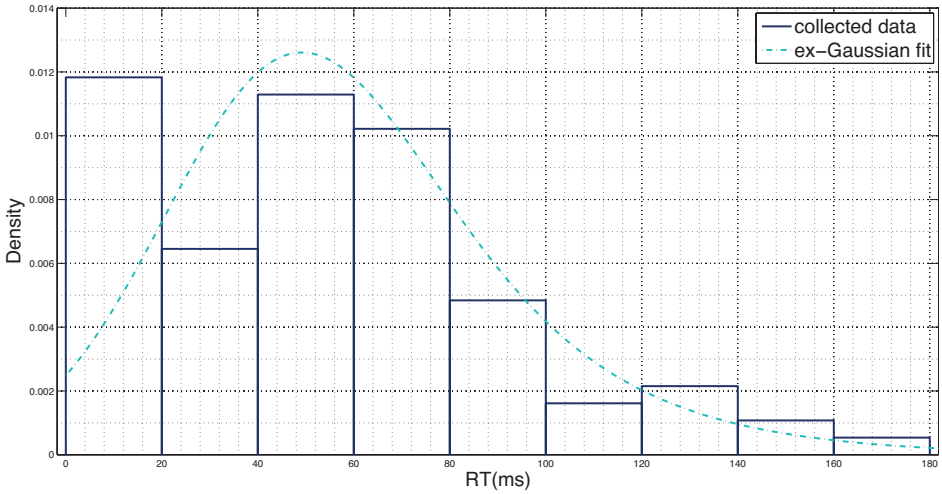


Figure B.3: Probability density function (PDF) of RT samples acquired in scenario B(4.3.3) using the 120fps recorded video

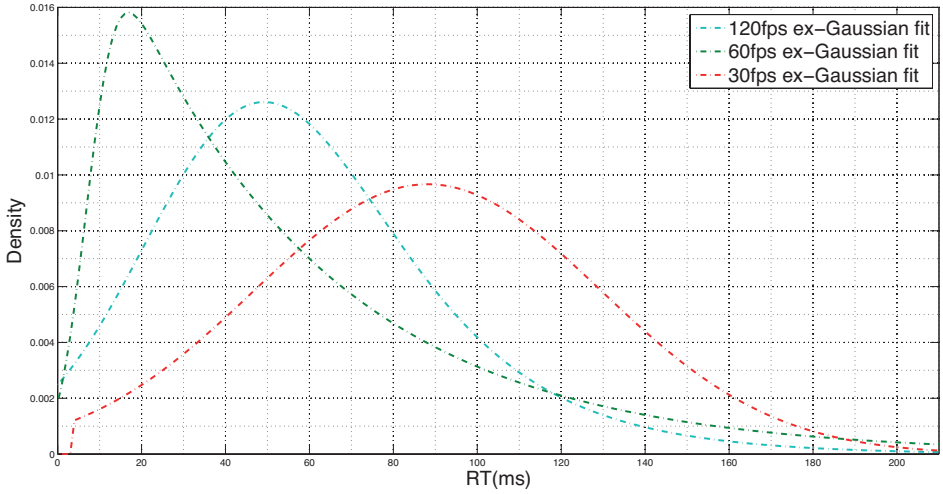


Figure B.4: All fitted Probability density functions (PDF) of scenario B(4.3.3)

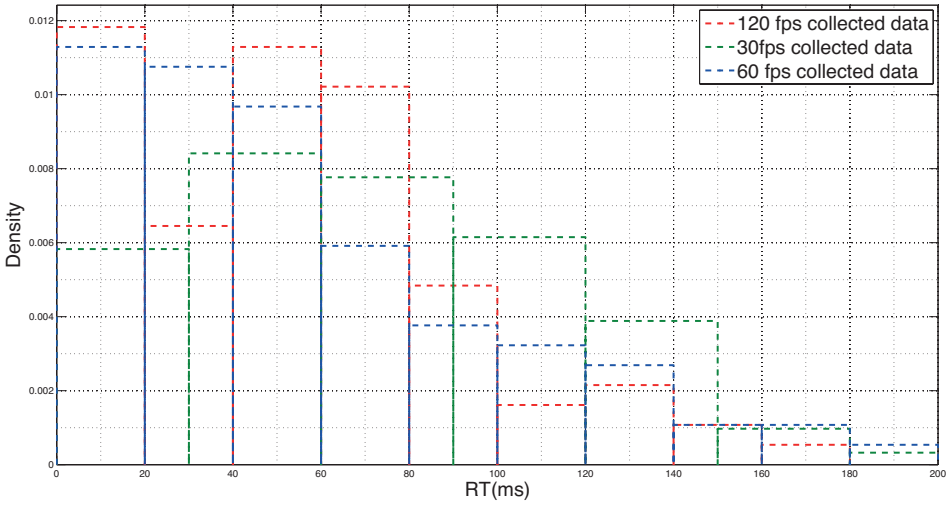


Figure B.5: Histogram of collected samples in scenario B(4.3.3)

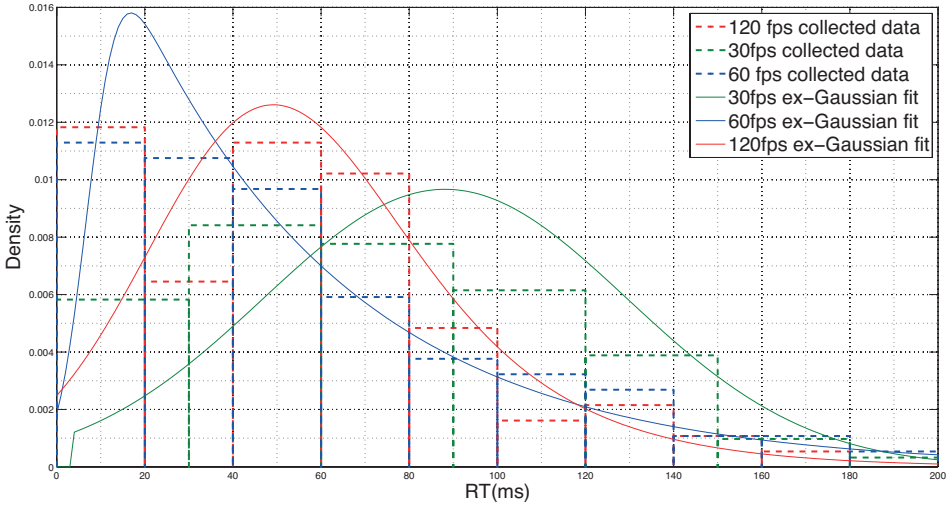


Figure B.6: Histograms and fitted PDFs in scenario B(4.3.3)

B.2 Results of Scenario C

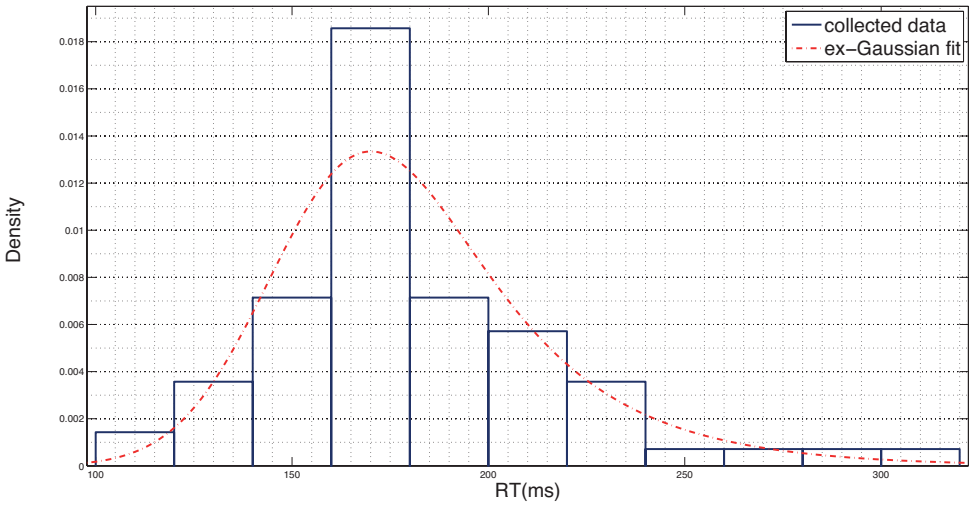


Figure B.7: Probability density function (PDF) of RT samples acquired in scenario C(4.3.3) using the 30fps recorded video

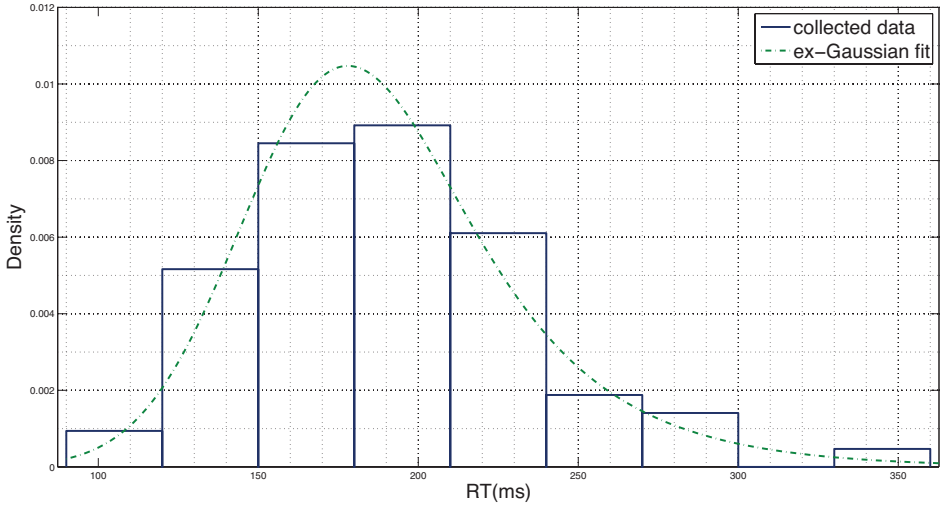


Figure B.8: Probability density function (PDF) of RT samples acquired in scenario C(4.3.3) using the 60fps recorded video

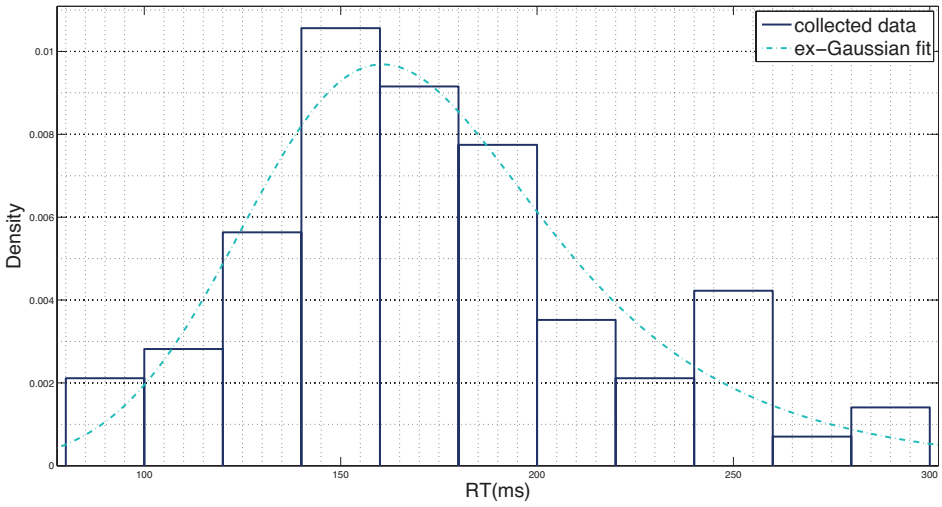


Figure B.9: Probability density function (PDF) of RT samples acquired in scenario C(4.3.3) using the 120fps recorded video

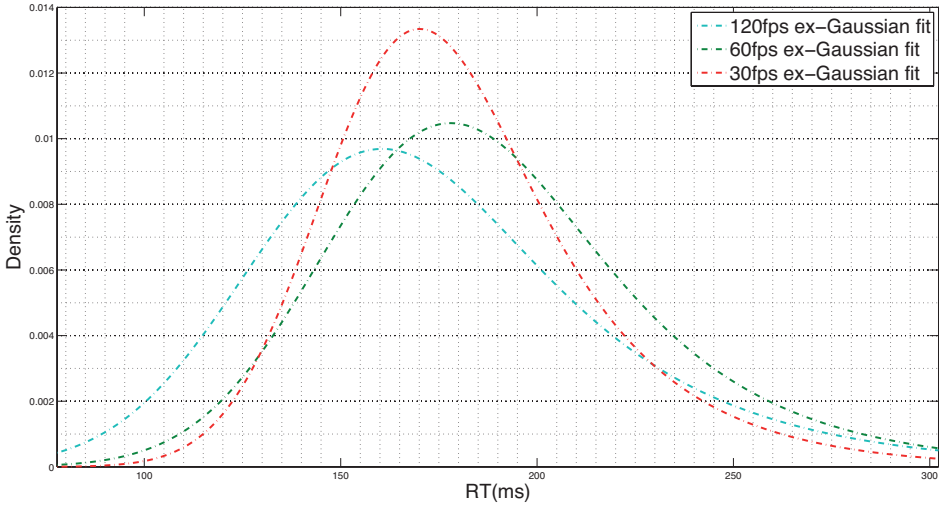


Figure B.10: All fitted Probability density functions (PDF) of scenario C(4.3.3)

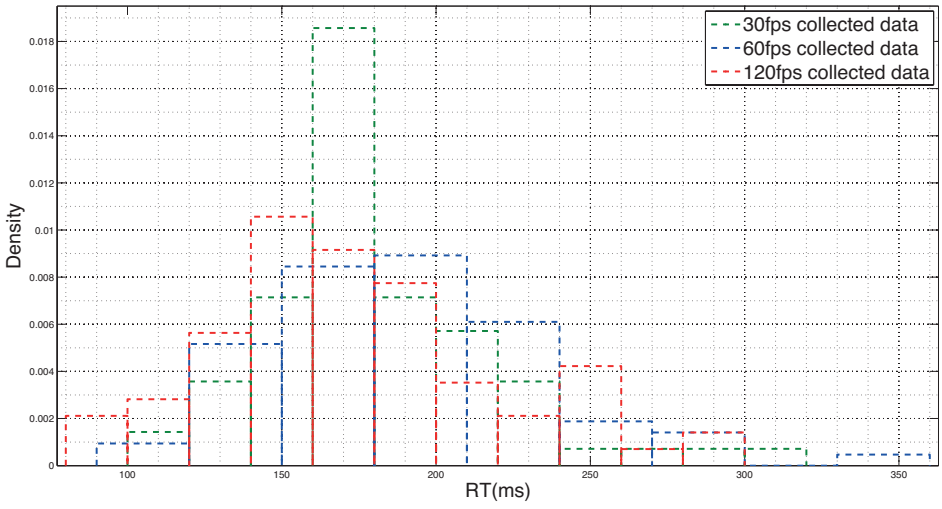


Figure B.11: Histogram of collected samples in scenario C(4.3.3)

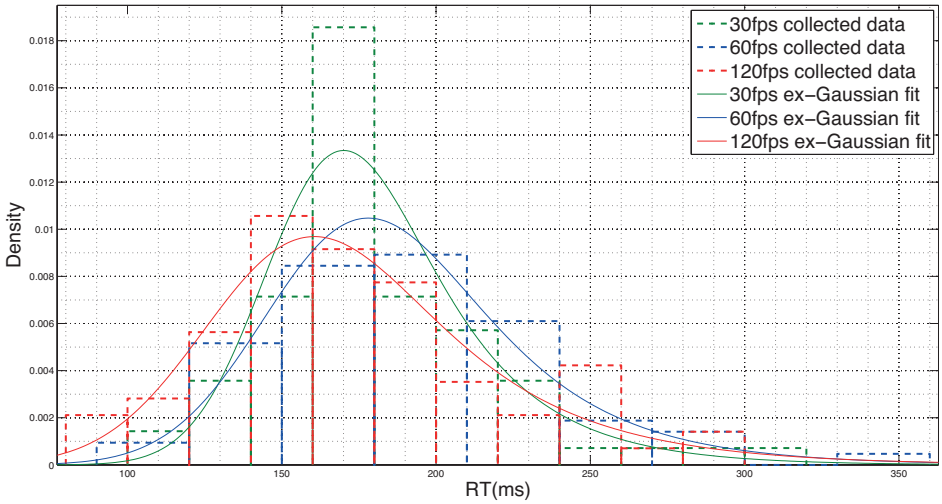


Figure B.12: Histograms and fitted PDFs in scenario C(4.3.3)

Appendix C

Depth cues

Monocular cues

Focus

When a scene is looked, the brain focuses each object that a person is paying attention building up a distance map. Therefore, in a still picture, the objects focused on are perceived nearby and the out of focused objects are distant.

Perspective

When looking at a scene, the relative size of the objects and the convergence of parallel straight lines into one point help to perceive and to build a relative distance map.

Occlusion

Objects at the front of a scene block the sight of the objects at the back. That is a clue about relative distance.

Lighting and shading

The light that is reflected in the surfaces of the objects and the shadows that appear are interpreted by the brain building a map of the position of the objects and the shape of these issues.

Colour intensity and contrast

The object that is further away has a reduced contrast and colour intensity. Therefore, with the different levels a relative distance map can be built.

Relative movement or motion parallax

Close objects appear to be moving faster than distant objects when an observer moves.

Vergence

When the eyes have to focus an object depending on how far away this object is, the eyes will be pointing this object almost in parallel or making up an angle. Thus, the position of the eyes or vergence aids our brain to figure out the perceived depth.

Binocular cues

Stereopsis

The stereopsis is the difference between the two images perceived by each eye. Furthermore, it is possible to calculate the depth according to these differences.

Appendix D

Proposal Implementation

D.1 Delaying video using a standalone device

This is an outline for a future implementation because this task is out of the scope of this thesis. It is focused on the DVI-D specification (see Group (1999)) due to the fact that it is an extended interface and most of the equipment used in this project is capable to manage it.

As it can be appreciated in figure D.1, the proposal would be made up of four blocks. The transmitter and receiver would be responsible to encode and decode Transition-minimized differential signaling (TMDS) signals respectively. These blocks are already implemented by several manufacturers and there are integrated circuits that perform this function. Therefore, the new implementation would have to deal with streams signal (see figure D.2).

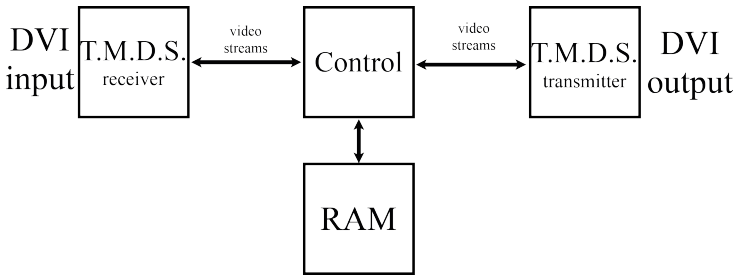


Figure D.1: Proposed implementation

These streams are composed of four control signal [CTLX], two signals to define when a line (horizontal sync [HSync]) and a frame (vertical sync [VSync]) have been transmitted or received, the colour information signals [RED, GREEN, BLUE] with eight bits per each one, data enabled [DE] signal and the clock signal.

Hence, the Control block would be responsible for storing the data signal and forwarding again, depending on how many delay would want to be introduced. This

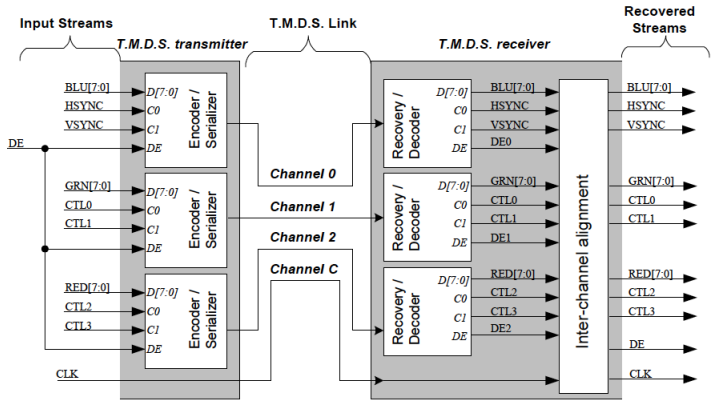


Figure D.2: Single link T.M.D.S. Channel Map (Group, 1999)

value could be introduced using a keyboard implemented in the same device or developing a communication link with a computer to set the configuration through a user interface. That decision depends on the developer. In this last case, the link with the computer would have just a transfering function. The operating system must not to be able to reach the hardware of this device.

Appendix E

Adding delay

E.1 Delaying audio

The used tool was `pacat` which corresponds to sound server called PulseAudio. It is used by most recent Linux distributions.

`pacat` is a simple tool for playing back or capturing raw or encoded audio files on a PulseAudio sound server.

The script line used is:

```
sudo schedtool -R -p 99 -e ./audioDelayPipeline.sh 1 1
```

```
#!/bin/bash
#$1 $2 are parameters
pacat -r --latency-msec=$1 -d alsa_input.pci-0000_00_1b.0.analog-
stereo | pacat -p --latency-msec=$2 -d alsa_output.pci-0000_00_1b
.0.analog-stereo
```

`-r | --record`

Capture audio data and write it to the specified file or to STDOUT if none is specified.

`--latency-msec=MSEC`

Explicitly configure the latency, with a time specified in miliseconds.

`-d | --device=SINKORSOURCE`

Specify the symbolic name of the sink/source to play/record this stream on/from.

`-p | --playback`

Read audio data from the specified file or STDIN if none is specified, and play it back.

E.2 Delaying video

Delaying video was achieved using a capture card and running a pipeline script corresponding to GStreamer(`http://gstreamer.freedesktop.org/data/doc/gstreamer/head/manual/html/index.html`) which is a pipeline-based open source multimedia framework.

It is a powerful framework for creating a wide range of streaming-media applications, to manipulate audio or video or both. The pipeline-design allows to perform several functions just piping some commands.

The script used was:

```
sudo schedtool -R -p 99 -e ./gst-videodelay-pipeline.sh 0
```

```
#!/bin/bash
#$1 refers to amount of delay in seconds
gst-launch-0.10 -v -m decklinksrc mode=17 connection=1 !
ffmpegcolorspace ! frei0r-filter-delay0r delaytime=$1 !
ffmpegcolorspace ! xvimagesink sync=false
```

Appendix F

Results of interviews

F.1 Interactive rehearsal

F.1.1 Survey

Q1i. Was the camera in front of the screen disturbing at beginning of test?

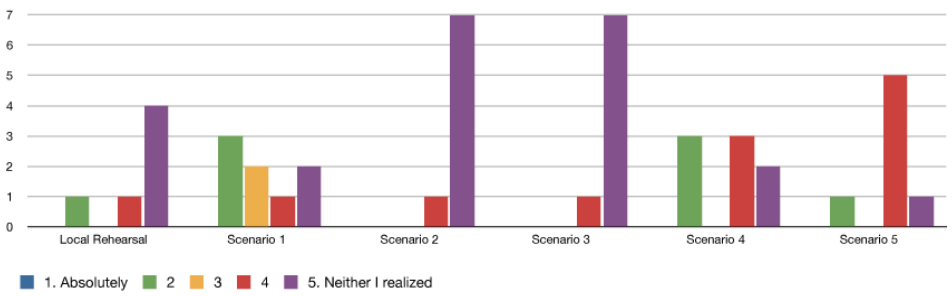


Figure F.1: Results of question 1 in interactive rehearsal

Q2i. Quality of the image

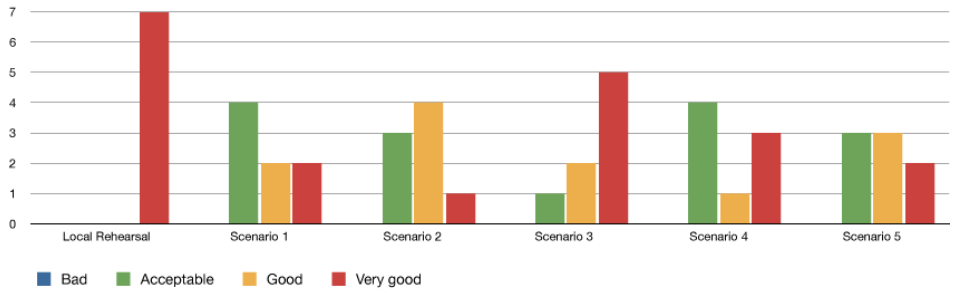


Figure F.2: Results of question 2 in interactive rehearsal

Q3i. ”Distance” between the conductor and the musicians

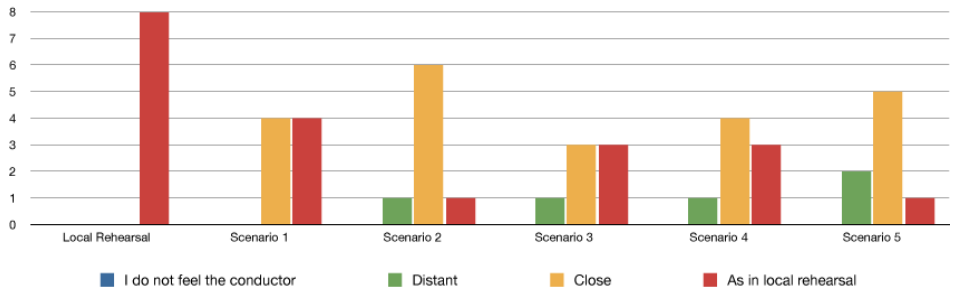


Figure F.3: Results of question 3 in interactive rehearsal

Q4i. Quality of the ”eyes to eyes” communication between the conductor and musicians

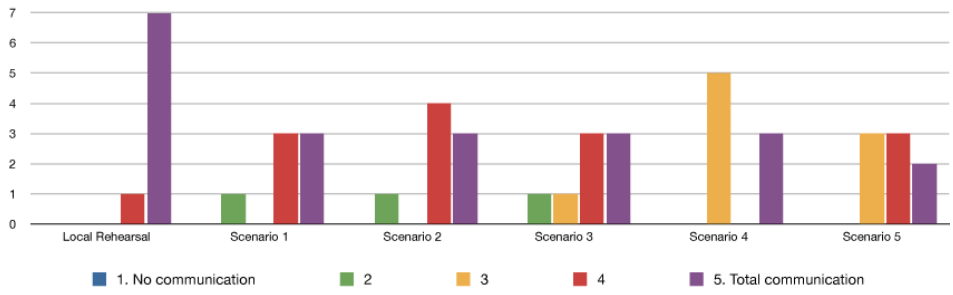


Figure F.4: Results of question 4 in interactive rehearsal

Q5i. Are you looking at the conductor's hands ?

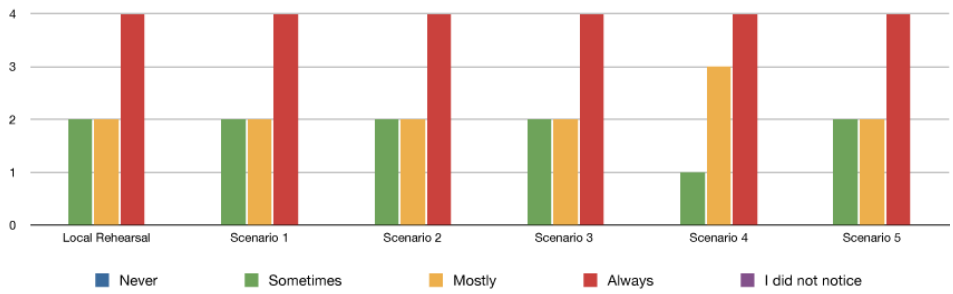


Figure F.5: Results of question 5 in interactive rehearsal

Q6i.Are you looking at the conductor's face ?

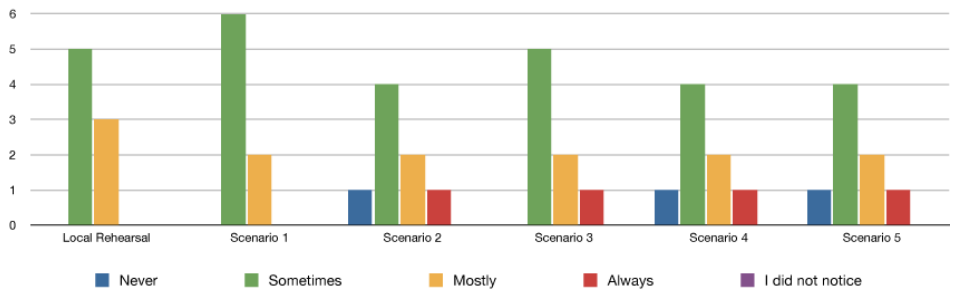


Figure F.6: Results of question 6 in interactive rehearsal

Q7i. Quality of dynamics

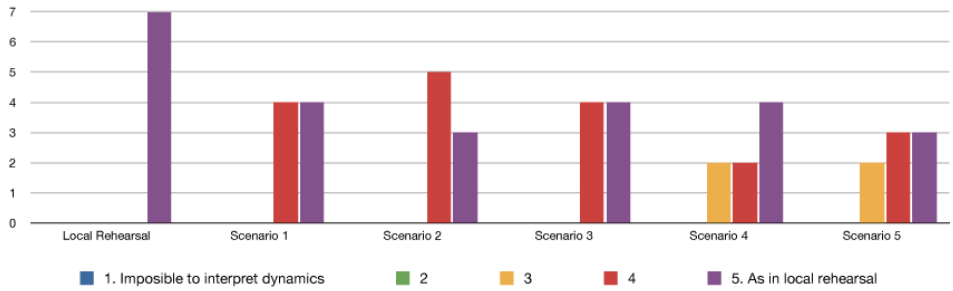


Figure F.7: Results of question 7 in interactive rehearsal

Q8i. Ability for phrasing/musical expression

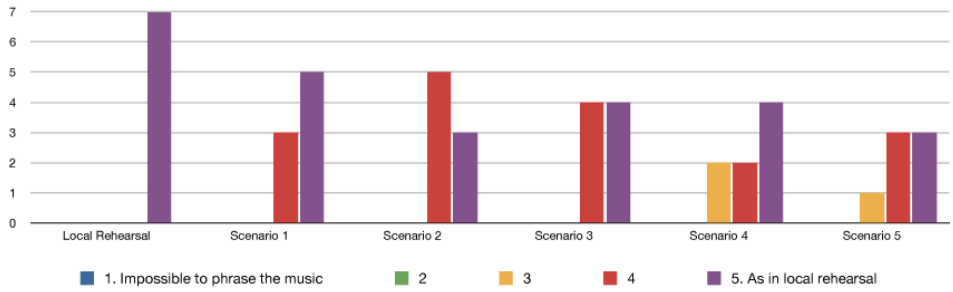


Figure F.8: Results of question 8 in interactive rehearsal

Q9i. Ability to find the ”tempo giusto”

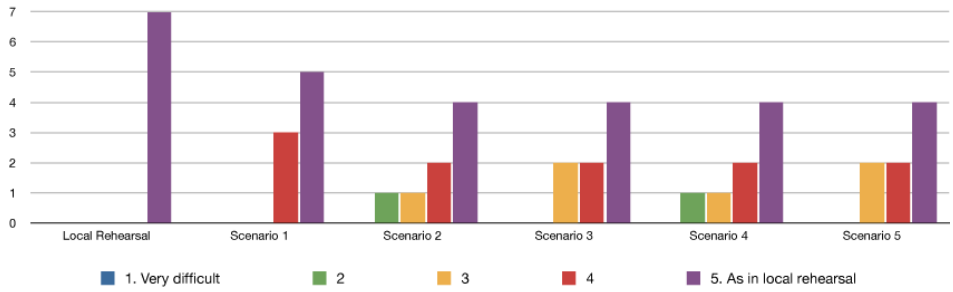


Figure F.9: Results of question 9 in interactive rehearsal

Q10i. Was the camera in front of the screen disturbing at the end of tests?

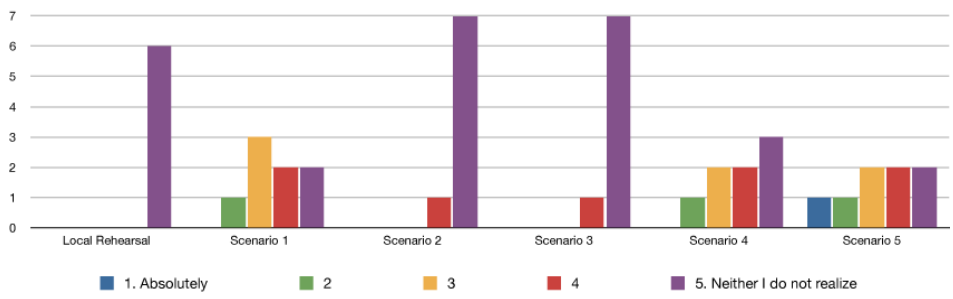


Figure F.10: Results of question 10 in interactive rehearsal

F.1.2 Participants' comments

Local Rehearsal 1

- "*Mange spm. irrelevant, da dirigent var tilstede*".
- "conductor in same room (local rehearsal)"

Scenario 1

- "I was very aware that this was a different situation and was less relaxed than in regular rehearsals".
- "Flickering image. Can make us tried in the long run?"
- "Conductor's clapping (3rd test) directly behind camera"
- "*Flimring på bildet, litt dårlig lys, hendene til dirigenten av og til utenfor bildet. Lydkvalitet og dynamikk/frasering var irrelevant i denne testen*".

Scenario 2

- "*Kamera ble flyttet så det ikke var foran bildet. Dynamikk/frasering, lyd var irrelevant*".
- "Did not feel the conductors' presence. Might as well have been a recording".
- "Local camera was moved down since test2, out of sight line. Actually no sound from conductor due to increased distance from her microphone".
- "The distinction between looking at the face and looking at the hands is maybe not very useful. Personally, I try not to focus too much on the details of the body language of the conductor, but rather to perceive the totality".

Scenario 3

- "Better image quality and lightning. Makes it easier".

- "I find it more necessary to look at the conductor's hands all the time while clapping together with her during this scenario".

Scenario 4

- "*Kameraet var nå hevet igjen, og dekket den ene hånden til dirigenten*".
- "Clapping was a bit more difficult".
- "Camera was located in height with the conductor's hands, which was very disturbing/annoying".
- "Local camera was moved up, almost in front of conductor, especially when clapping. Otherwise no noticeable difference from previous tests".
- "During this scenario I could see that the conductor really had to concentrate - especially during the clapping".

Scenario 5

- "*Kameraet dekket den ene hånda. Lyd og dynamikk/frasering var irrelevant*".
- "Camera covers conductor's hands".
- "The conductor did not seem to need to concentrate as much during this scenario than during the last one".
- "Local camera almost in front of conductor's hands, especially when clapping. Tired/low concentration due to room temperature + ambient light made image contrast apparently lower. Otherwise as previous tests. (No sound from conductor due to distance from her microphone)".

F.2 Videos interview

F.2.1 Survey

Q2v. Quality of the image

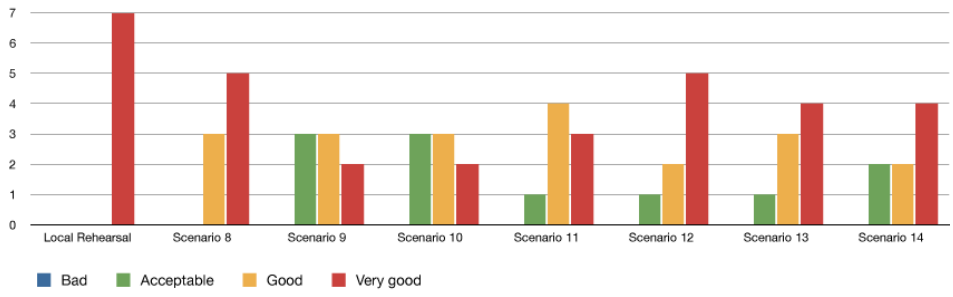


Figure F.11: Results of question 1 conducting according videos

Q3v. ”Distance” between the conductor and the musicians

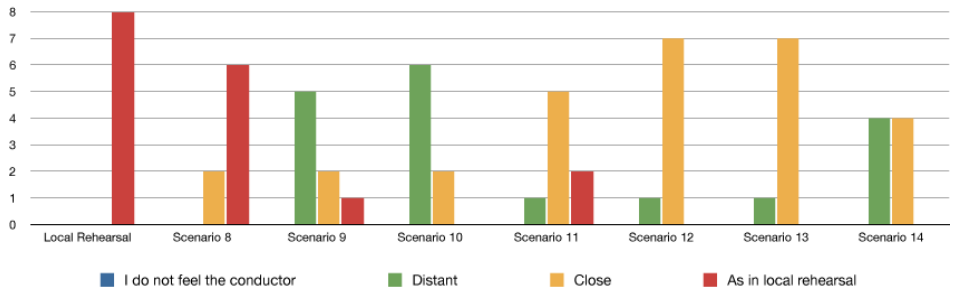


Figure F.12: Results of question 2 conducting according videos

Q4v. Quality of the ”eyes to eyes” communication between the conductor and musicians

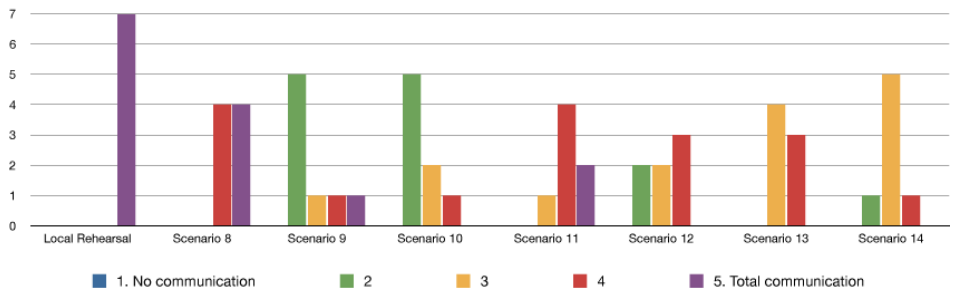


Figure F.13: Results of question 3 conducting according videos

Q5v. Are you looking at the conductor’s hands ?

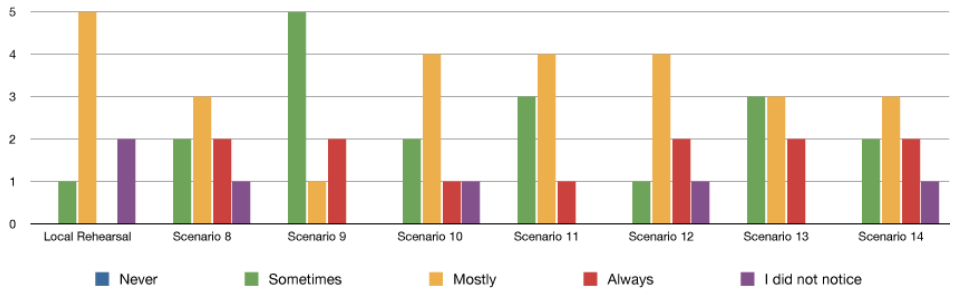


Figure F.14: Results of question 4 conducting according videos

Q6v.Are you looking at the conductor's face ?

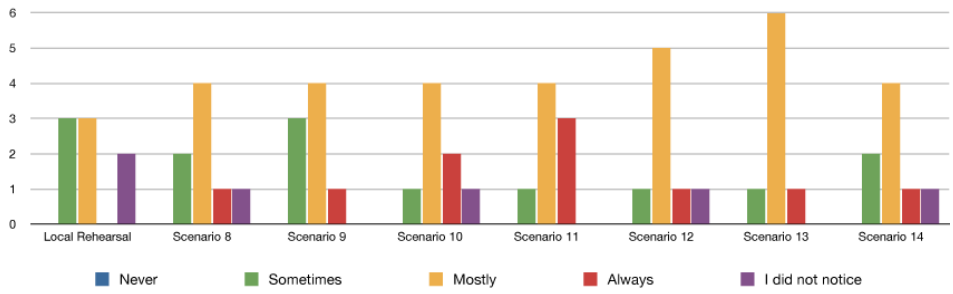


Figure F.15: Results of question 5 conducting according videos

Q7v. Quality of dynamics

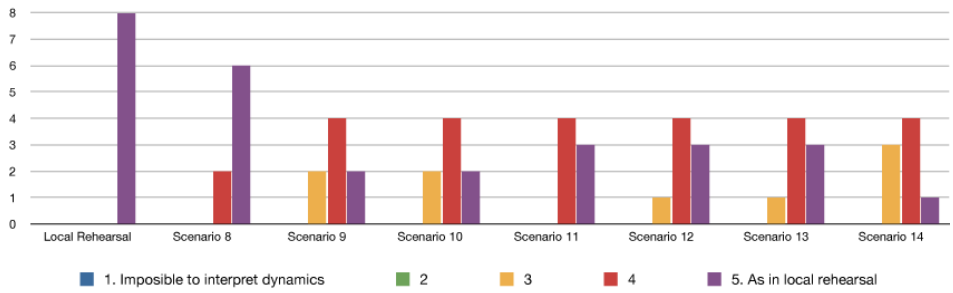


Figure F.16: Results of question 6 conducting according videos

Q8v. Ability for phrasing/musical expression

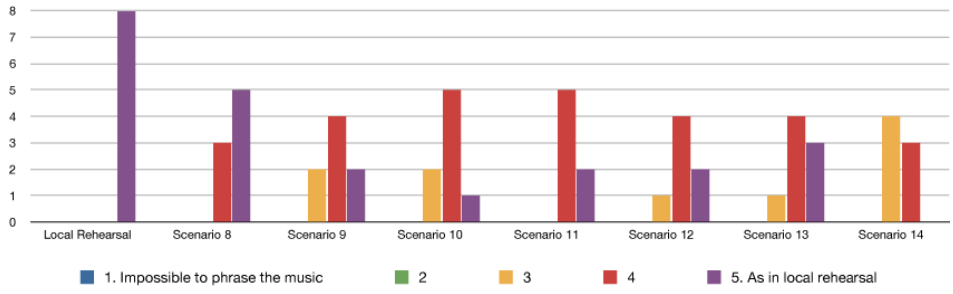


Figure F.17: Results of question 7 conducting according videos

Q9v. Ability to find the "tempo giusto"

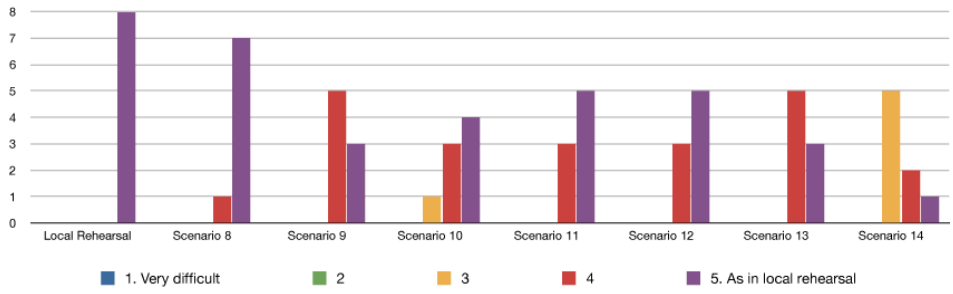


Figure F.18: Results of question 8 conducting according videos

F.2.2 Participants' comments

Local rehearsal 2

- "*Mye irrelevant, da dirigent var tilstede*".
- "Local rehearsal, no camera/image".

Scenario 8

- "*Stort, godt bilde!*"
- "Interesting difference from basic tests (clapping etc); now Grieg performance: I looked more at conductor's face, less at hands. Plus: did not notice camera on front now".

Scenario 9

- "*Kameraspm. irrelevant. Litt lugging i bildet, men akseptabelt. Merket at dirigent ikke hadde kor foran seg. Dette ga ""død"" kommunikasjon, men selve dirigeringen fungerte*".
- "No camera in front. Main annoyance: not smooth image / movements. Could also notice that conductor was "off-line", i. e. without direct feedback from choir".
- "*Det var tydelig at dette var et opptak av dirigenten, siden det ikke var noen " direktekommunikasjon" mellom det som faktisk foregikk i koret musikalsk sett og de grepene dirigenten gjorde. Det ble mekanisk*".
- "It is easy to see that she was conducting without a choir on front of her face is more ""dead"", and the eyes looks without focus, with more distance".
- "White background made it much easier to see what she was doing. I noticed that there was no choir, as there was no response to what we did. But the dynamics, phrasing was a lot clearer than in live conducting. In that respect it was easier to follow and the result sounded more interesting to me".
- "It was evident that the conductor was alone in the room, with no singers to communicate with. She was looking at us without the focus she would have were we in the same room".

Scenario 10

- "*Det var bedre når dirigenten denne gangen så i retning av de ulike stemmegruppene, men det var fremdeles tydelig at dette var et opptak og at det ikke var noen direkte kommunikasjon mellom dirigent og kor*".
- "*Kameraspm. irrelevant. Dirigent så rundt på et fiktivt kor og virket noe mindre ""død"" i blikket, men det fungerte ikke godt når hun så på min stemmegruppe. Det føles aldri som om hun ser på meg annet enn når hun ser rett i kameraet*".
- "Contrast between her clothes and background made perception easier. Her looking in different directions made it seem more real. Her hands seemed somewhat slower, I could almost divide the movement inwards in separate parts as compared to the live version. -It was strange to feel that she wanted and really tried to communicate with the ensemble, - but did not succeed at all!".
- "The gaze of the conductor is not focused on the singers. The addressing if the various groups (SATB) works only to a degree - our positions on the room did not correspond enough with the location she assumed we had. This may cause confusion and may thus not work as intended - in worst case can it be counter-productive".
- "No camera. Conductor had far too wide conception of imaginary choir. Looked as if she tried to avoid looking at any of us. This is natural, since eye-to-eye contact can be obtained with image even when oblique viewing angle to screen, if only the image person looks into the camera. This was not the case here".

Scenario 11

- "I felt like she was looking at me".
- "I really felt that the conductor was communicating with me! I felt that she was looking at me during the whole song".
- "*Det ble mer personlig nåsom dirigenten så i kamera, siden det føltes som om dirigenten så direkte på meg. Dette var bedre enn begge de tidligere videoene*".
- "The eyecontact/focus was better than in scenario 9".
- "*Litt mindre død i blikket enn nr. 9, men merket fortsatt godt at det var dirigert uten et kor til stede. Dette gir mindre innlevelse fra meg, og dårligere fraserings- og dynamikkformidling*".
- "No camera in front. Compared to test 9 (first with Grieg): very near same quality, perhaps smoother images now? Main problem, due to off-line (recorded conducting): no feedback to choir on our performance, conductor looked very indifferent to what we produced :-(".

Scenario 12

- "I could feel that he was not looking at us. I happen to know this conductor and I know he is like that in person as well!"
- "*Noe mindre dødt blikk, men tror det kan skyldes at det var helt ny og fremmed dirigent, så vi opplevde ikke kontrasten til når han er til stede. At vi var ukjent med ham gjorde også at vi fulgte ham lit dårligere enn Vivianne, uavhengig av at det var på film*".
- "With a new conductor we are not familiar with, the hands draw more focus, as we are not accustomed to his movements and way of conducting".
- "*Det var tydelig at han forestilte seg et sittende kor, siden blikket hans hele tiden søkte lavere enn hvor høyt vi står i virkeligheten. Om han hadde henvendt seg på samme måte men samtidig sett i kamera på en eller annen måte hadde hadde et vert veldig godt*".
- "Both Michael and Vivianne are very experienced conductors. It is important that they are able to imagine the music, and that they know from experience which gestures they shall use to get the desired effects".
- "New, unknown conductor.Did not look far to the side, almost as if he knew the width of the choir. Quite easy to follow. Also some facial expression that could make me believe that he actually heard us. Probably ok due to unknown person, better than I would expect from unknown".

Scenario 13

- "She is an excellent communicator and managed to communicate through the camera. The movement of her arms was a bit "split up" by the video".
- "dynamics and shifts in tempo was hampered by the quality of the recording. this problem was evident with Grete, who made more shifts i tempo than the other conductors".

- "*Grete var veldig til stede i musikken selv uten kor, og jeg oppfattet kommunikasjonen bedre enn de andre opptakene. Hun dro tempoet mer opp og ned, og da merktes det at det var vanskeligere å følge godt. Tror det kan skyldes at hun ikke hørte oss og kunne justere. Det er merkbart mye lettere å dirigere til opptak når tempoet er jevnt og rolig*".
- "Since the conductor feels a little too introvert (conducting with closed eyes), it was more difficoult to feel the connection with her than the two others. Even her hands and figures are very clear and easy to read".
- "*Ukjent dirigent med litt overraskende tolking. Men til å være første gangen gikk det rimelig bra*".
- "Another new conductor. Easy to follow, also nearly the impression that she ""heard"" the choir. But seemed a little less concentrated than the Canadian conductor, hence a little more ""distance"" to the choir".
- "Again, this is a very experienced musician, with a good imagination. She even seems to imagine some common problematic musical elements".

Scenario 14

- "Her hands were blurry. Harder to follow tempo and onsets this time".
- "*Det er tydelig at det ikke finst noen mulighet for dirigenten å justere koret i tilfelle feilskjær eller misforståelser slik man ville kunnet i virkeligheten. Dette forringer opplevelsen av direktekommunikasjon*".
- "*Dette var eneste klipp hvor det ikke ble slått full slagfigur til enhver tid, og det gjorde at koret sakset internt, og med dirigent. Dermed ble det alt i alt vanskeligere å synge. Jeg antar at når dirigenten ikke får feedback er det ekstra viktig med rytmisk tydelighet så behovet for justering aldri oppstår*".
- "I've given the questions Ability for phrasing/musical expression and Ability to find the ""tempo giusto"" a relatively low score. The reason is that the conductor didn't conduct so clear as she did in the last scenario. She tried to free herself from the strict conduction /figures - which I felt did not succeed".
- "Conductor seemed more inspired/committed than previous test (or maybe I have got more used to her). Rather easy to follow, although some problems (unclear conducting) in the middle, not related to technique/video".
- "This time, the conductor did some strange things that made it difficult to follow her. Maybe the video was too slow to capture her rapid hand movements".
- "Hard to follow her tempo - probably because we are not familiar with her way of conducting. And as she gets no auditory feedback from the choir, she cannot adjust this. works probably best with a conductor the singers know well".

Appendix G

Permission contracts

NTNU
Norwegian University of
Science and Technology

Permission is granted to Daniel Puig Conca and Leif Arne Rønningen in order to publish the results of performed tests during May at Department of Telemactics at Norwegian University of Science and Technology (NTNU).

Participants:

Andrés Cervantes

Diego Salvador

Elaheh Vahidian

NTNU
Norwegian University of
Science and Technology

Permission is granted to Daniel Puig Conca, Leif Arne Rønningen and Vivianne Sydnes in order to publish the results of performed tests on the 25th of May 2012 at Department of Telemactics at Norwegian University of Science and Technology (NTNU).

Participants:

Anita Brevik

Anne Sigrid Imsen

Solveig Meland

Mariel Eikeset Koren

Terje Aandalen

Jon Bang

Martin Eikeset Koren

Lars Sydnes