Technical Sciences
Brassersplein 2
2612 CT Delft
P.O. Box 5050
2600 GB Delft
The Netherlands

www.tno.nl

T +31 88 866 70 00
F +31 88 866 70 57
infodesk@tno.nl

TNO report

RA 35515

Inter-destination Media Synchronization for TV broadcasts

Date 18 May 2011

Author(s) Rufael Mekuria

Supervisors Dr. M. Oskar van Deventer, Prof.Dr. Rob Kooij, Dr. Fernando Kuipers

Project name Graduate project - Inter-destination Media Synchronization
Key Words Television, Social TV, Synchronization, Measurements, Tool

Summary
This thesis presents a study on the application of inter-destination synchronization for TV-broadcasting. Inter-destination media synchronization implies synchronizing media output at different receivers. This thesis starts by investigating differences in media output between receivers of TV broadcasts at different locations and different technologies. To do this a measurement scheme is developed using media mining techniques and the fact that differences were found to be relatively fixed between receivers. Using this tool differences ranging from 0-5s were found depending on the technology and the channel used. The second aim was to test the user experience of inter-destination synchronization in (Interactive)-TV applications. After studying related social TV literature, a user test for the specific effect in social TV was developed. The test was performed at the KU Leuven using a test-panel of 36 users. The results show that contrary to the state of understanding of social TV, the social experience has little dependency on inter-destination synchronization in the 0 to 4s range. A soccer watching experiment was performed to investigate the experience of inter-destination media synchronization when an audio link is present. The thresholds found, of when the play-out difference becomes annoying or perceptible, was comparable to the social TV use case.

Rufael Mekuria:

MSc. Thesis: Inter-destination media synchronization for TV broadcasts

12-4-2011

Program: Telecom Engineering TU Delft Department:
Network architectures and services

Supervisors: Dr.Oskar van Deventer (TNO-ICT),
Dr.Fernando Kuipers (TUD)

Committee: Dr. David Geerts (KU Leuven),

prof. Robert Kooij (TNO-ICT)

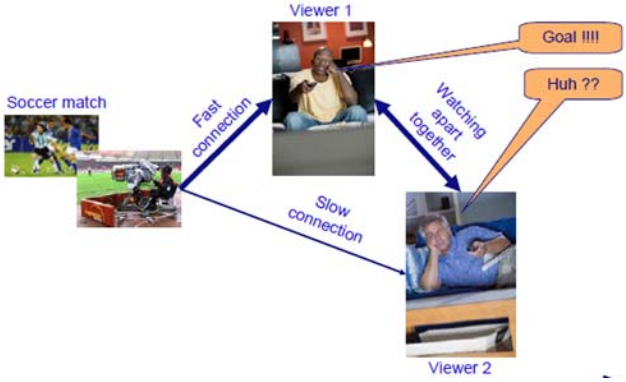
Outline of this presentation

- Introduction to Social TV and Inter-destination media synchronization
- Research Outline
- Inter-destination media synchronization measurement study
- Inter-destination media synchronization User experience study
- Conclusions/Future work
- Questions/Answers session

Social TV Example

From: Social TV Review by MIT Relying on relationships to rebuild TV audiences[1]

Social TV example: IDMS needed

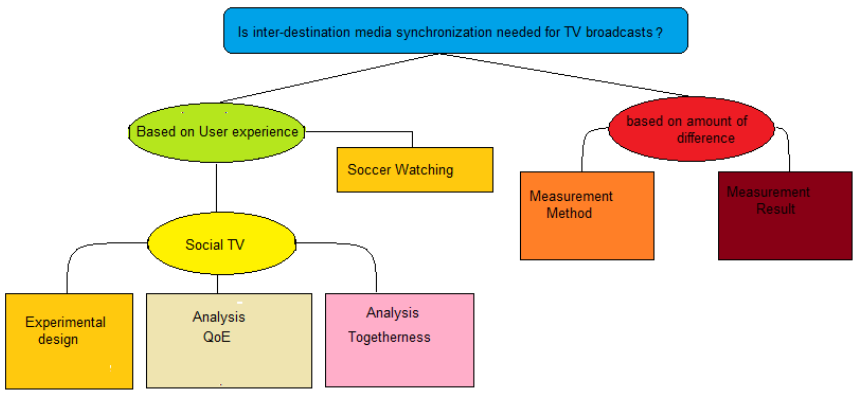


From: TNO- International ETSI standard for Inter-destination NGNLab.eu 2 Nov 2010

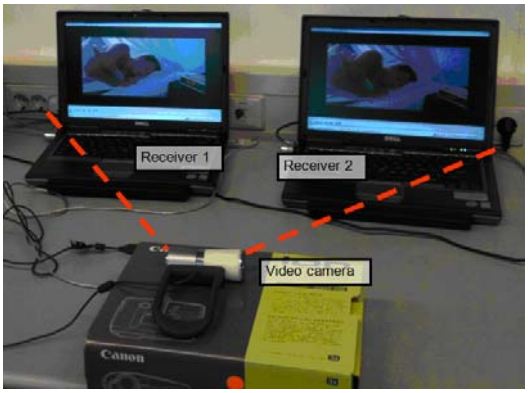
Research Question

- When is inter-destination media synchronization useful in TV broadcasting services?

Structure of Research Project

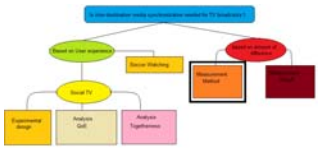


Physical lab setup

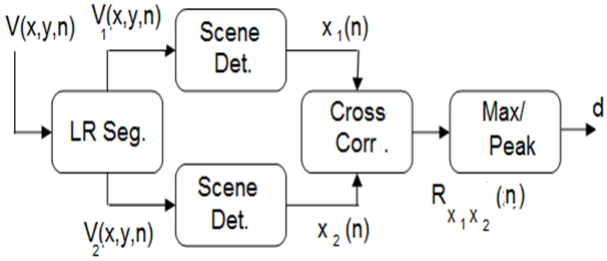


Advantages:
1.Broadly applicable

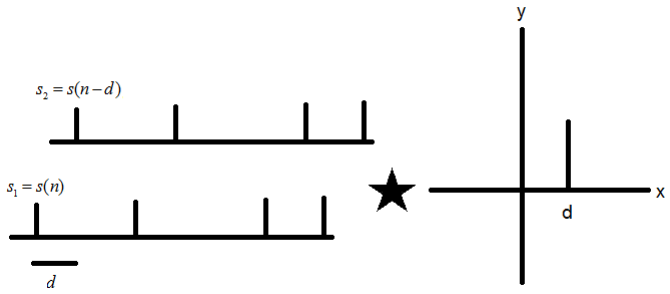
2. takes TV/set-top box delays into account
3. Mobile setup

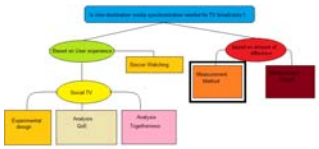


System outline

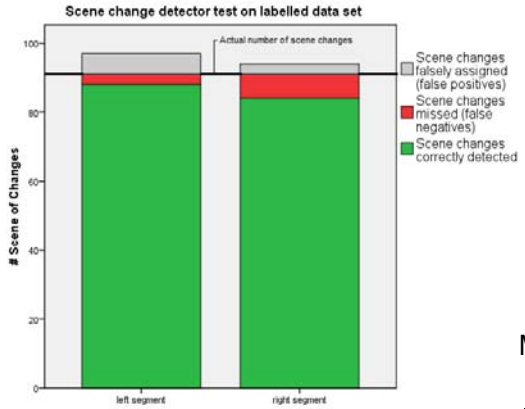


Principle of operation





Scene Change detector Evaluation



Pd= 94.5%

Pf=0.14%.

More than sufficient for our measurement purpose!

Applications Measurement Tool

- New QoS metric for TV, distributors can compare lag to competitors relevant to soccer watching
- Useful for companies that need measurements to synchronize interactive applications (games,ratings) to TV content
- Input lag relevant to gamers which has been hard to measure [4]
- Validation of synchronization solutions (how well do they work ?)

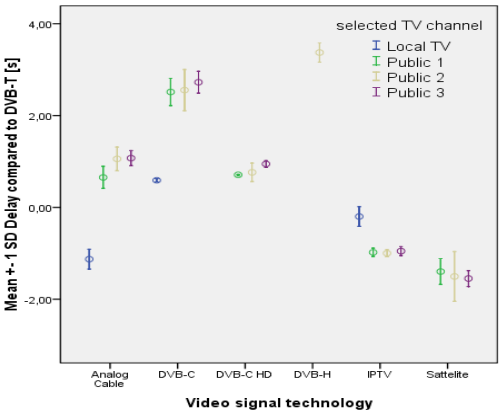
An addition to[5]

- Tuning of synchronization algorithms (performed in our user experiment)
- Previously no broadly applicable measurement system was available. System will be presented as a demo at the euroITV 2011 conference Lissabon (june 29-July 1)

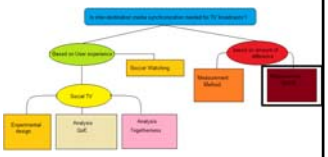
Measurement study TV broadcasting:

- Find how large differences are that occur and how much they change
- Den Haag, Zoetermeer, Delft,Leidschendam
- Measure broadcast TV (DVB-C,DVB-C HD,DVB-H,DVB-S, cable, DVB-T, IPTV), web TV and delays in set-top box.
- Done by comparing to DVB-T to other TV broadcasts (297 measurements)
- DVB-T was measured as a good reference
- Comparing web streams on different computers

Results Broadcast TV(manual)



Results web-TV (nos.nl)



	HD/HD	***HD/SD***	***SD/SD***
Silverlight	*0-2s*	*2-8s*	*0-2s*
Media player	*2-4s*	*4-8s*	*0-2s*

Conclusions

- For TV approximately fixed differences in small geographic area ranging from 0-5s
  depending on the technology/provider and TV channel
- Web differences up to 8s or perhaps even more are observed and change when the browser is restarted.
- Trial on delays in TV and set-top box did not show significantly large differences encountered compared to station and technology (0-200ms instead of 0-5s)

User Experience of IDMS in Social TV

- Social TV/Soccer
- We asessed both use cases in user trials
- KU Leuven August 2010 (36 couples) Quiz
- TNO januari 2011 (5 Couples) Soccer

User Experience assesment

- Adapting QoE metric defined by the ITU P.800 adapted to IDMS gave the questions shown below
- We measure Togetherness (Main benefit of Social TV according to previous research) based on IDMS (KU Leuven provided Togetherness Questionaire)

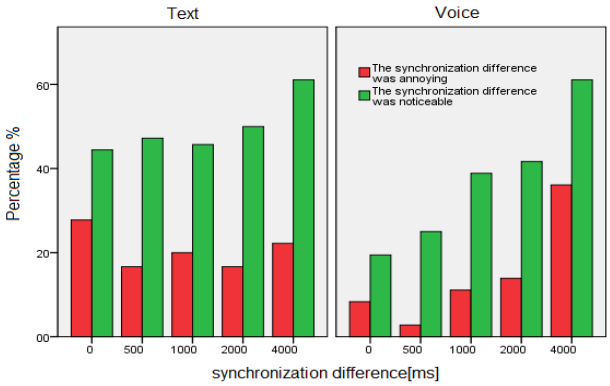
MOS Value	Impairment
5	the synchronicity difference is not perceptible
4	the synchronicity difference is perceptible but not annoying
3	the synchronicity difference is perceptible and slightly annoying
2	the synchronicity perceptible and annoying
1	the synchronicity difference is perceptible and very annoying

KU Leuven user trial August 2010 experimental design

- 36 couples 5 times text 5 times voice
- Randomized synchronization conditions
- Sociable genre,couples
- Synchronization method CWI validated within 40ms accuracy
- Experimental design and user test proved succesful

High likeability, togetherness and chat activity observed

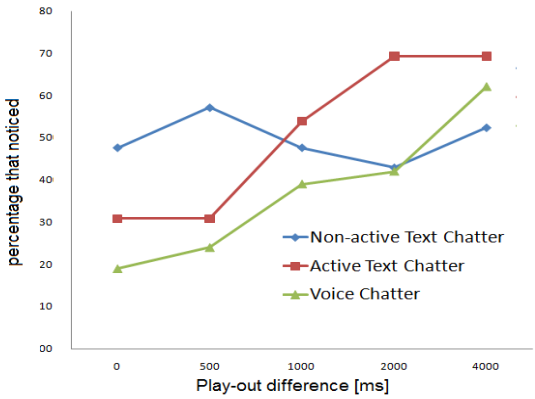
Results perception and annoyance



Statistical Analysis Perception and Annoyance

- Psychometric method not used due to lack of data, spread in data and high variance between participants characteristic of social TV
- Non parametric Cochrane´s Q test used instead
  - takes participant(within) variance into account
  - Assumes no distribution on the data
  - Works for binary response variable
- Each condition is compared to the synchronized condition and considered significant if $p<0.05$
- Voice chatters notice 1s or more, text chatters 4s or more
- Voice chatters get annoyed at 4s
- Text chatters don´t notice or get annoyed significantly compared to being synchronized

Noticeability based on IDMS



Togetherness

- Togetherness was obtained from averaging responses to six questions on a 1 to 7 scale
- The questions were consistently answered (cronbach's alfa and gutmann's split half both above 0.8 which implies consistency in social research)
- Voice chatters were statistically tested to feel more together than text chatters in a paired samples t test $t(179)=-7{,}143$ $p<0.001$
- The play-out differences(IDMS) were statistically shown not to have a meaningful effect on togetherness
- Active chatters (>400) were also tested to feel more together than non-active chatters $t(178)=-6{,}2$ $p<0.001$

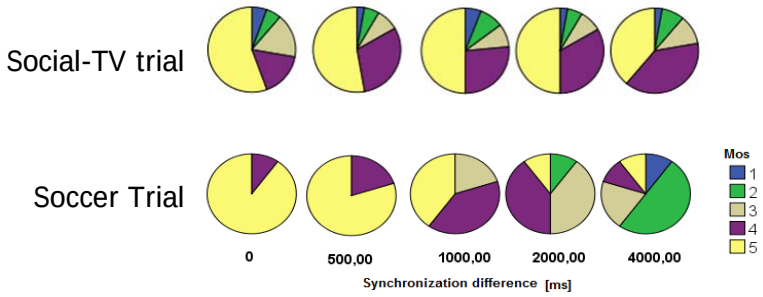
Conclusion

- Active chatters and voice chatters both feel more together and notice differences in an approximately similar way. Non active text chatters notice less but feel less together.
- Therefore we recommend play-out difference maximum of 1s such that active text and voice chatters obtain a seamless social TV experience.

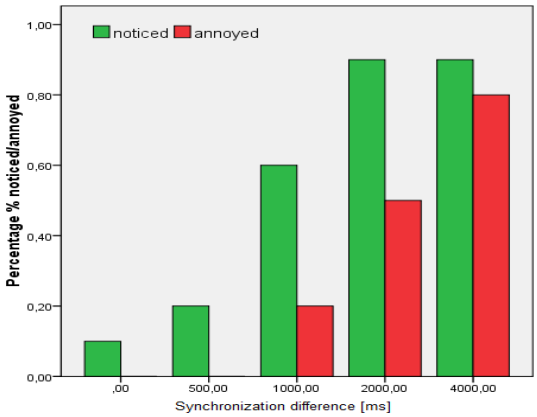
Soccer Watching Experiment

- The aim is to check the effect of IDMS in broadcasting in the case of two co-located soccer match viewers(audio link)
- This represents both the neighbours shouting and a social TV situation

Results Soccer QoE watching vs trial 1



Perception and annoyance



Useful inter-destination media synchronization for (Social) TV

- In the internet between receiver clients of web streams(0-8s)
- In P2P services were play-out differences have been shown to range up to 6s [11]
- Between given TV technology and channel play-out is approximately fixed, otherwise between 0 and 5s
- Play-out differences less than 1s are sufficient for social TV and soccer watching
- Togetherness/Social Aspect of IDMS smaller than expected which contradicts previous research[2]

Comparison to Video conferencing

- Traditional application of inter-destination media synchronization is in Video conferencing
- Often used well managed IP networks (assuming no transcoding) do not introduce delays significant for social TV (< 1s) [12]

	Video Conferencing	**TV Broadcasts**
Network requiring synchronization	Same network/protocol	Different Networks/protocols
seamless experience	0-200ms range	0-1s range
Typical differences encountered	0-1 s range	0-6 s range

Questions ?

References(1)

- [1] Bulekey MIT technology review 2010 R10 Social TV http://www.technologyreview.com/communications/25084/
- [2]. **Shamma, Bastea-Forte, Joubert, Liu.** Florence : ACM CHI EA '08 CHI '08 extended abstracts on Human
- *Enhancing online personal connections through synchronized sharing of online video.*
- [3]. **Clipsync.** www.clipsync.com. [Online] 9 3 2011. clipsync.com.
- [4]. 20. **posts, various.** LCD input lag/video delay measurement (PC and console games responsiveness). *AVS forum.* [Online] 27 8 2007-2008. [Citaat van: 9 3 2007.] http://www.avforums.com/forums/lcd-led-lcd-tvs/612503-lcd-input-lag-video-delay-measurement-pc-console-games-responsiveness.html.
- [5] **Nunome, Tasaka.** *Application level QoS comparison of inter-destination media synchronization schemes for continuous media multicasting.* sl : IEICE Transactions on communications vol. 87, 2004.
- [6]. http://radio-tv-nederland.nl/dvbt/digitenne-kpntv.html
- [7]. **Geerts, Cesar, Bulterman.** *The implications of program genres for the design of social television systems.* San Francisco : ACM, 2008.

References(2)

- [8]11. **Oehlberg, Duchenaut, Thornton.** *Social TV: Designing for distributed, sociable television viewing.* Athens : Euro ITV 2006, 2006.
- 12. *Enhancing online personal connections through synchronized sharing of online video.*
- [9]. **Shrimpton-Smith, Bieke Zaman, Geerts.** *Coupling the users: The benefits of paired user testing for iDTV.* Athens : Ablex Pub. Corp. International Journal of Human-computer Interaction vol:24 issue:2
- [10]. **Brinkman, W.P.** *Handbook of Mobile Technology Research Methods.* Delft : Nova science methods, 2009.
- [11]. **Fallica, Lu, Kuipers, Kooij, Van Mieghem.** *On the quality of experience of SopCast.* Cardiff : IEEE Next generation mobile applications conference, 2008.
- [12.] **ITU-T G.1050.** *Network model for evaluating multimedia transmission performance over internet protocol.* sl :ITU-T, 2007.

Thank you for your attention!

Social TV

- Top 10 of emerging technologies according to MIT Technology review 2010[1]
- First large scale test in user homes was performed by TNO-ICT in Enschede the Netherlands (ConnecTV) in 2007 where the concept was shown viable

Large Business Interests

SOCIALVISION

BBC iPlayer TV Radio

Yahoo! Zy

clipsync

Watchitoo

synch tube

BT

MTV MUSIC TELEVISION

NBC

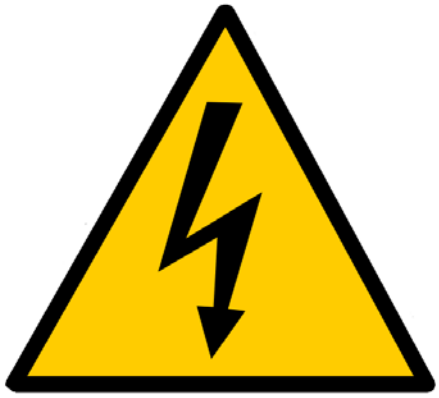
Research Question

- When is inter-destination media synchronization useful in TV broadcasting services?
  - 1 Based on how much play-out difference occurs between live broadcasting services
  - 2 Based on how differences affect the user experience (QoE) ?

This was never investigated before but crucial for motivating IDMS research or implementation!

Previous work [2] and industry [3] suggest inter-destination synchronization enhances the social experience. This thesis aims to answer this question precisely and extensively using two measurement studies

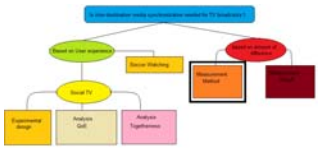
Live Demonstration



Short/Long term Merits of this research project

- IDMS implementation requirements for Social TV, a hot topic in industry
- Measurements useful for inter-destination media synchronization for interactive quiz/rating shows to TV
- Motivates standardization activities such as for ETSI TISPAN and ietf as performed by TNO
- New QoS parameter relevant to Soccer fans
- Better user experience, new applications services business and advertisement models
- Save the broadcast industry??

Applications Measurement Tool



- New QoS metric for TV, distributors can compare lag to competitors relevant to soccer watching
- Useful for companies that need measurements to synchronize interactive applications (games,ratings) to TV content
- Input lag relevant to gamers which has been hard to measure [4]
- Validation of synchronization solutions (how well do they work ?)

An addition to[5]

- Tuning of synchronization algorithms (performed in our user experiment)
- Previously no broadly applicable measurement system was available. System will be presented as a demo at the euroITV 2011 conference Lissabon (june 29-July 1)

Performance of measurement tool

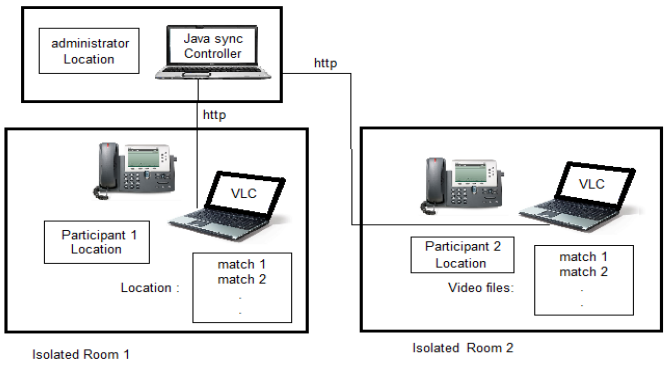
- 40ms abolute average difference found comparing with slightly inaccurate manual measurements (average 6ms)
- However when comparing two receivers DVB-T and cable or two cable receivers tool is very accurate and produces the same results consistently (0 for cable,cable, fixed value for DVB-T,cable)
- Difference mainly attributed to manual measurement error.
- Mathematical analysis suggest that assuming a constant scene change probability model and the scene detection parameters measured we are far above the minimum value, this makes the system robust to lighting changes/size changes which can decrease the detection probability.
- This is also clearly observed in practice

DVB-T as a reference Signal

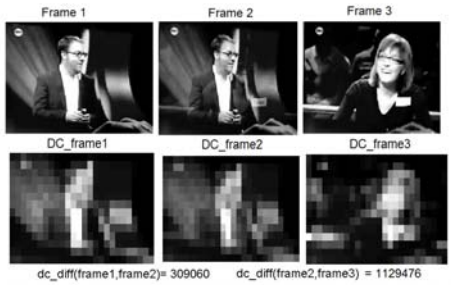
- The area tests consist of homes in Zoetermeer, Den Haag en Delft
- According to broadcast source [5] this area is covered by 5 DVB-T Transmitters operating at the same frequency
- A pilot set of 20 10 min measurements showed that signal strength does not effect play-out difference for DVB-T (only Quality)
- This makes us believe DVB-T is a good reference signal

For indirectly comparing play-out differences between broadcasts

Experimental Lab setup trial 2 at TNO

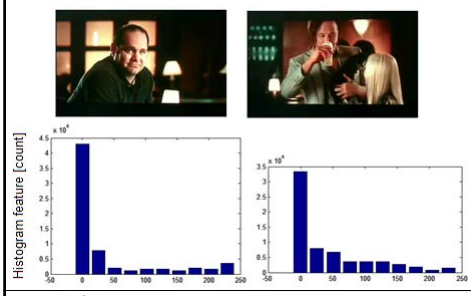


Scene Change detection features(1)

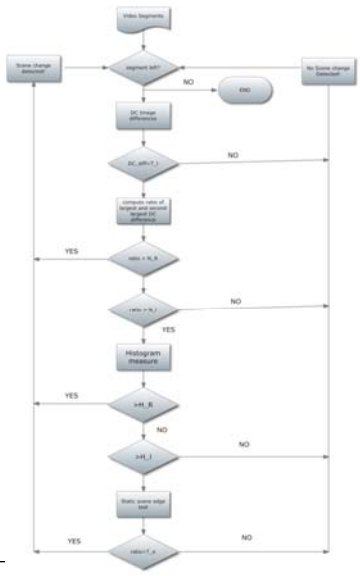


Play-out difference measurement study: Measurement Method (4)

Scene Change detection features(2)

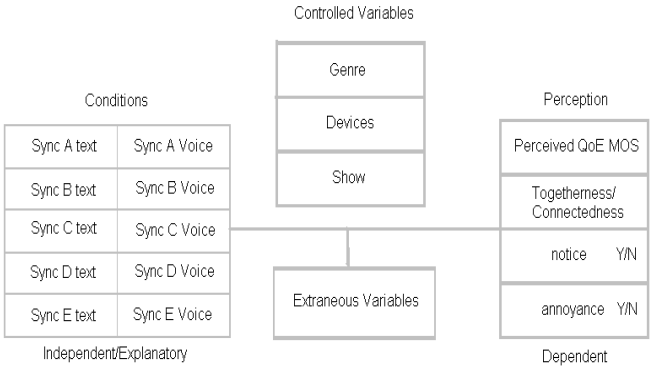


Play-out difference measurement study: Measurement Method (5)



Parameter	Value
N_R	4
N_L	1.5
T_low	1.5*segment_Area
H_I	2000
H_R	4000
T_e	0.175

Experimental Design



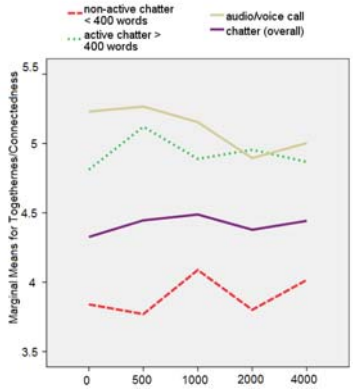
Controlled Variables

- Use case quiz of the pappenheimers was chosen(KU Leuven) as quizes give high sociability[4],are liked[4] and give reasonably constant amounts of content in each period.
- Laptops were used for the watching experience based on practical considerations
- Couples (friends,family or partners) were recruited which has been shown to enhance social interaction in Social TV[5]. Interaction is necessary to detect play-out differences as test the case without audio interference from the other TV [6]

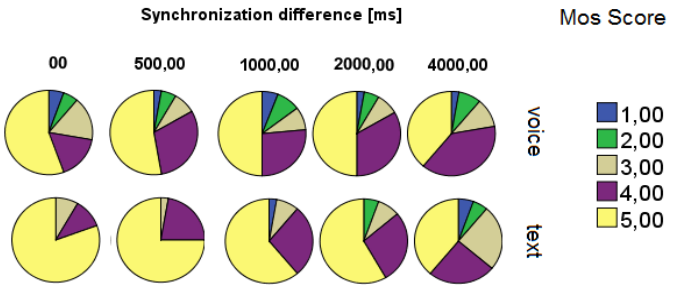
Validity Considerations

- List from [7] checked for applicability
- Randomized synchronization conditions to remove possible habituation/fatigue effects
- Clear instruction to prevent drop-outs
- Recruitment broad audience (no specific target groups) to obtain recruitment validity

Togetherness based on IDMS



Results QoE MOS



Other effects Perception and Annoyance

- Half of voice chatters saw the episode before, they noticed play/out differences best
- Voice chatters that did not see the show before noticed much less well, actually non significant similar to text chatters
- Active chatters, more than 400 words per session, approximately 10 per minute did notice differences of 2s and 4s significantly and therefore seem to behave more like voice chatters

TNO-ICT (now TNO)

- Not for proft research institute
- Social and interactive TV are two large focal points of TNO-ICT
- TNO-ICT performs research and standardization activites related to inter-destination media synchronization
- It is believed to enhance interactive and social TV which will become important in the future for TNO's clients
- www.tno.nl

Measurements on live web casts

- Nos.nl broadcasts news and politics
- Co-located PC's on the same network were tested
- Results vary heavily upon restarting the browser
- Play-out difference remains approximately constant when browsers are not closed
- 4 repetitions per case, range is reported as values vary heavily
- 0-8 second range depending on HD/SD or whether silverlight or mediaplayer is used

Delft University of Technology

Faculty of Electrical Engineering, Mathematics and Computer Science
Network Architectures and Services

Inter-destination media synchronization for TV broadcasts

R.N. Mekuria
1199234

Committee members:
Supervisor: Dr.ir. F.A. Kuipers
Mentor: Dr.ir. M.O. van Deventer
Prof.dr.ir. R. Kooij
Member: Dr. D. Geerts

April 12, 2011
M.Sc. Thesis No: PVM 2011 – 068

Preface

This thesis is the result of my master's study in Electrical Engineering at the Delft University of Technology. In my master's studies I obtained knowledge by attending courses in Telecommunications and Media and Knowledge engineering. Thanks to the kind suggestion of Fernando Kuipers I was able to join TNO-ICT for a master's thesis project. This thesis is also the concluding report of this project.

Firstly I have to express my gratitude to Oskar van Deventer. Oskar was always open for discussion, new ideas and eager to learn despite his very busy loaded schedule. Thank you for supporting me with my Thesis and in the TNO organization.

Secondly I would like to thank Fernando Kuipers for his supervision, advice on academic research and suggestions for improving this work.
My gratitude also goes out to Robert Kooij for his supervision at a slight distance, his motivation and the after work football matches.

I would also very much like to thank Ray van Brandenburg, who was always very friendly and very interested in my work. I hope we will someday have a chance to do a project together. Moreover I would like to thank the many people I have worked with at TNO and who have participated in the user experiments. Thank you! Without you this would not have been possible. I would also like to thank David Geerts from KU Leuven and Ishan Vaishnavi from the CWI for the times in Leuven we spent together, it made a large part of this thesis possible. I also very much appreciated the support of Dick van Smirren and Rixte Thomas during the Thesis work period. I also like to thank my friends for their support and friendship during this Thesis work.

Note: This Thesis contains a supplementary DVD containing deliverables to TNO-ICT. It contains the developed software, datasets from experiments and analysis, figures, videos containing measurement recordings, (Co-) authored international publications and a supplement survey of statistical techniques used

Summary

This thesis presents a study on the application of inter-destination synchronization for TV-broadcasting. Inter-destination media synchronization implies synchronizing media output at different receivers. This thesis starts by investigating differences in media output between receivers of TV broadcasts at different locations and different technologies. To do this a measurement scheme is developed using media mining techniques and the fact that differences were found to be relatively fixed between receivers. Using this tool differences ranging from 0-5s were found depending on the technology and the channel used. The second aim was to test the user experience of inter-destination synchronization in (Interactive)-TV applications. After studying related social TV literature, a user test for the specific effect in social TV was developed. The test was performed at the KU Leuven using a test-panel of 36 users. The results show that contrary to the state of understanding of social TV, the social experience has little dependency on inter-destination synchronization in the 0 to 4s range. A soccer watching experiment was performed to investigate the experience of inter-destination media synchronization when an audio link is present. The thresholds found, of when the play-out difference becomes annoying or perceptible, was comparable to the social TV use case.

List of Symbols and Abbreviations

ACR	Absolute Category Rating
ANOVA	Analysis of Variance
CSV	Comma Seperated File
DCR	Degradation Category Rating
DCT	Discrete Cosine Transform
dh	Histogram difference
$Diff_{dc}$	DC Image difference
DipLib	Digital Image Processing Library
DVB	Digital Video broadcasting
DVB-C	Cable Digital Video Broadcasting
DVB-H	Digital Video Broadcasting for Handheld
DVB-S	Satellie digital Video Broadcasting
DVB-T	Terestrial Digital video broadcasting
ETSI TISPAN	Telecommunications and Internet converged Services and Protocols for Advanced Networking
F_{dc}	DC image
$H_A(i)$	Histogram of frame A
HCI	Human computer Interaction
IDMS	Inter-destination media synchronization
IPTV	Internet protocol Television
ITU	International Telecommunications standardization organization of the United Nations
MATLAB	Matrix Laboratory
MOS	Mean Opinion Score
NTIA	National Telecommunications and Information Administration
p_d	detection probability
p_{f_p}	false positive probability
$P_{sc,}$	Scene change probability
QoE	Quality of Experience
QoS	Quality of Service
RR	Receiver Report
RTCP	Real Time Transport Control Protocol
RTP	Real Time Transport Protocol
Rxy(n)	cross-correlation function between y and x
s(n)	Original scene changes
SIP	Session intiation protocol
SPSS	Statistical Package for Social Sciences
SR	Sender Report
TNO	Nederlandse organisatie voor toegepast natuurwetenschappelijk onderzoek
VLC	Video LAN media player
VQM	Video Quality Metric
x(n)	Detected scene change function
XR	Extended Report
Y(n),X(n)	Sequence of Video Frames
α	Static scene change coefficient

Table of Contents

1. Introduction

1.1 Background Information

The project is carried out at TNO Information and communication technology which is an applied research institute in the Netherlands. Its mission is to bring innovation and technology to small and large businesses. The expertise of the institute is broad and includes various technical topics in telecommunication, usability issues and application development. One of TNO's focus points is in interactive television and social TV. TNO was the first in the world to perform a large scale field trial of social TV [1]. TNO also contributes to standardization of interactive and mobile TV by actively contributing to European and International standards such as ETSI TISPAN. A particular focus of TNO for enhancing interactive TV services is inter-destination media synchronization. TNO has several patents in this technology and has made contributions to ETSI TISPAN to enable inter-destination media synchronization in IPTV [2]. However, not much information is available about the effect of inter-destination media synchronization on the user experience in interactive applications such as social TV. Also no information is available on how much play-out difference occurs between television broadcasts at different locations. To improve their knowledge about these two aspects TNO put the author on an MSc project for the duration of 9 months.

1.2 Research Purpose

The main problem faced by TNO-ICT engineers was to obtain play-out difference data between broadcast channels and about the user experience of inter-destination media synchronization in various applications. Measuring differences play-out and user experience is not trivial as is shown in this thesis. The main research purpose of this thesis is to collect as much information as possible about play-out difference and user-experience. The information is needed to support the next generation of social and interactive TV and obtain inter-destination synchronization recommendations for such applications. These interactive and social TV applications will allow the broadcasting industry to deploy new business and advertisement models around TV content and provide an enhanced user experience.

1.3 Related Work

There is little publicly available data of play-out differences between receivers of TV content. Most studies only focus on play-out differences caused in a single technology during the transmission, regardless of the application. For example in the area of inter-destination media synchronization simulation studies in networks with similarity to the internet were performed in [3] and [4] . A model for delay impairments encountered in IP based networks based on actual ISP data is given in [5]. However in TV broadcasting practice many different technologies exist such as the different DVB technologies described in [6] and analogue cable. This makes the results of these experiments only partially relevant.

Many subjective perceptual experiments to test the QoE of synchronization have been performed. For example already in 1996 the QoE aspects of audio-video synchronization and jitter were studied experimentally in [7].

The QoE of inter destination synchronization was also studied experimentally for two different video conferencing applications in [8] and [9]. Guidelines and considerations for social TV which is one of the main applications under consideration where given in [10] and [11]. In both studies inter destination synchronization was not explicitly given as a main factor that enhances the user experience. A study [12] claims that inter-destination synchronization improves the shared TV watching experience.

1.4 Focus of the Thesis

The main focus of this thesis is to investigate the usefulness of applying inter-destination media synchronization for broadcast TV. First we aim to measure play-out difference between receivers of unsynchronized TV broadcast in various conditions (technology, location, TV), large differences will imply that inter-destination media synchronization is likely to be useful, small differences indicate otherwise. Secondly we investigate the user experience of inter-destination media synchronization with more rigor than in [12] where only some random user responses were used with a main focus on social TV. If large effects on the user experience are found, inter-destination media synchronization is likely to be useful, in the case of small differences it is not.

The complexity of the research in this thesis lies in practical aspects of measuring play-out difference and the user experience and statistically interpreting the results. Measuring play-out differences has been shown to be difficult in various applications. Examples include accessing input lags of TV's, validation of inter-destination media synchronization solutions and the measurement study performed in this thesis. Measuring play-out difference between receivers is even more difficult in the measurement study performed here, because we want to measure in different proprietary (closed) networks with different technologies. To solve this problem we present a robust and broadly applicable measurement method.

What makes measuring the user experience of inter-destination media synchronization difficult compared to for example inter stream synchronization (audio-video) synchronization is that the role of social factors and genre tends to be bigger. Also no ITU standards for this specific use case exist. In this thesis a test is developed to measure the effect of inter-destination synchronization that takes these factors into account. Also considerable attention is paid to the appropriate statistical analysis of such a user test.

Taking the results of both experiments into account we aim to develop general guidelines and recommendations for applying inter-destination synchronization for TV broadcasts.

1.5 Thesis Outline

In chapter 2 we develop a robust measurement tool for measuring play-out difference (inter-destination media synchronization). This constitutes a concrete tool for measuring the quality of inter-destination media synchronization on the application layer. This chapter solves not only the measurement problem faced by TNO but also a general class of measurement problems on playback devices.

In chapter 3 this tool is used to do measurements on various broadcasting systems ranging from the traditional and DVB broadcasting technologies to web streams. A small pilot measurement between set-top boxes and TV's is performed. This chapter gives a picture of the inter-destination media synchronization quality of broadcasts in a small geographical area in the Netherlands. The most important play-out difference values encountered are given. Also the property how they vary in time is given attention. It shows that broadcast-TV and web based TV have different synchronization properties and requirements.

In chapter 4 an experiment to test the effect of inter-destination media synchronization on the user experience is developed. The test developed aims to isolate the pure effect caused by inter-destination media synchronization. To do this relevant academic social TV literature is studied and methods from social research are employed.

In chapters 5 and 6 the results of the test which was performed in KU Leuven Belgium together with dr. David Geerts are analyzed for their merits. Chapter 5 gives the thresholds of when the difference will become annoying or noticeable and the MOS indicators. Chapter 6 analyzes the effect of inter-destination synchronization on the social experience. Chapter 7 presents a small user experiment that tests the effect of inter-destination media synchronization when watching football matches when an audio link is present.

Chapter 8 gives the conclusions, which are general guidelines for applying inter-destination media synchronization for broadcast TV. In the future work section in chapter 8 we also highlight some other potentially useful applications of inter-destination synchronization in broadcast TV.

2. A new method for measuring inter-destination media synchronization for research purpose

This chapter proposes a new method to measure inter-destination media synchronization (play-out difference). By researching related measurement methods and their applicability in section 2.1 we found that measuring inter-destination media synchronization for TV is not trivial as different applications use different methods. The system developed in this chapter on the contrary is broadly applicable.

The system will be employed in chapter 3 to measure inter-destination synchronization in television broadcasts. Later in chapters 4 and 6 we use it to validate two different synchronization solutions for a user test.

In section 2.2 we define the requirements of a measurement system. The system design, implementation and demonstration are presented in sections 2.3, 2.4 and 2.5 respectively. In section 2.6 the system performance is assessed. Section 2.7 defines different use-cases for the developed tool.

2.1 Existing methods for measuring play-out difference

In this section we analyze existing available techniques that can be deployed to measure play-out differences. The main aim of the system should be that it can compare play-out difference between different TV-broadcasts and that it can be used to validate synchronization solutions.

2.1.1 Timestamp Methods for Computing play-out differences

Inter-destination media synchronization was traditionally mainly applied in video conferencing and multi point video communication systems. Measurement methods for comparing play-out between receivers were often based on comparison of time stamps. A good example is the often cited study comparing the performance of inter-destination synchronization algorithms [4] . The study from [4] used timestamps and a small fixed delay estimate per terminal to measure application level inter-destination media synchronization (e.g. play-out difference) in a simulation setup.

If headers of packetized multimedia in video conferencing or IPTV contain a timestamp with the time the packet was sent, play-out difference can be estimated by comparing this “sent time” to the actual reception time. This measures transmission delay of the packet which is often assumed approximately equal to the play-out difference.
In modern video conferencing and IPTV standards RTP [13] is often used for packetized multimedia transmission. RTCP [13] is often used for service information and feedback. These protocols are given as an example to illustrate how play-out difference can be estimated from timestamps. The right side of Figure 2.1 shows the structure of an RTP packet which consists of a UDP and IP packet, the RTP header adds information on timing, sequence number and payload type to allow the systems to keep track of synchronization and transmission characteristics. The left side of Figure 2.1 shows a simplified process of a receiver terminal setting up a connection using session initiation protocol SIP [14] to receive a video and audio stream which can be a live video conference or IPTV. Upon confirmation by the server the receiver receives RTP video/audio with headers containing the sequence numbers and the timestamps with the information when the packet was sent. During the RTP transmission RTCP control information is exchanged periodically in parallel. The sender sends sender reports (SR’s) and the receiver sends receiver reports (RR’s) which both contain summaries of transmission-reception information. By comparing the actual arrival time of the packet to the RTP timestamp (sent time) the receiver can keep track of the transmission delays encountered. If access to the receiver is possible this data can be a used for measurement purposes.

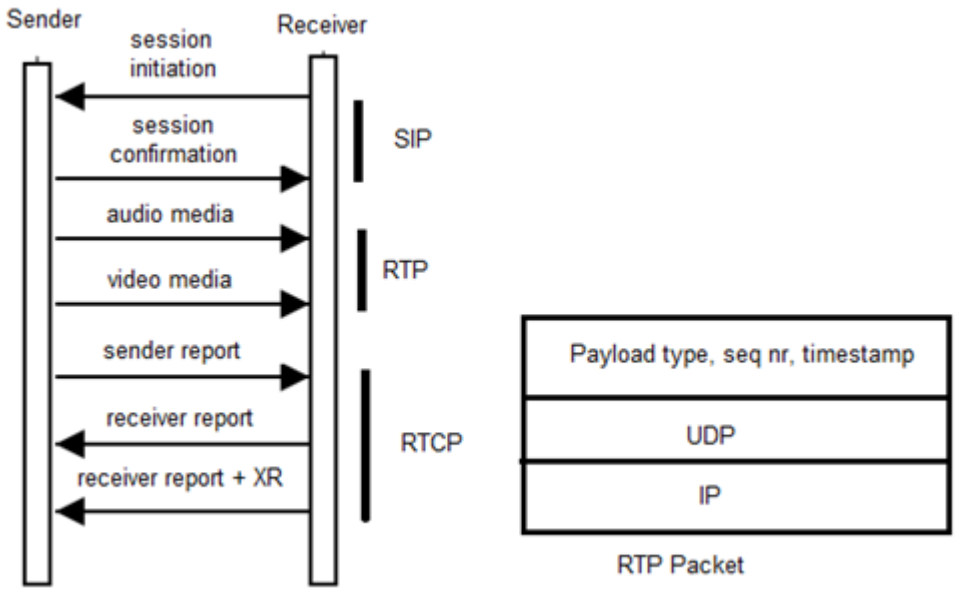


Figure 2.1 Inter-destination information in the RTP/RTCP protocol suite

However measurement at the server is often preferred for control purposes.
Originally the RTCP receiver report feedback did not contain fields to store the packet received/presented times. However RTCP has an XR extension where this information can be added. Recently an XR block extension containing this information was proposed [15]. By sending the arrival times back to the video server, play-out differences can be tracked. The proposed XR block from [15] contains fields for both the packet presented and the packet received times.

2.2.2. Input lag measurements/clock display measurements

Measuring play-out difference between two terminals has recently become relevant to players of video games. Players of TV-based video games started to notice delays introduced by digital TV's. These delays are caused by image processing routines such as scaling and enhancement. These delays can spoil the gaming experience. The effect has been reported in gaming magazines [16] and [17]. Independent research performed in [18], [19] and [20] showed that HDTV lags vary between 30 and 90 ms depending on the television type of signal used.

Figure 2.2 Play-out difference measurement of input lags

As it can be seen it is not trivial to read out the numbers on the TV and laptop screen. These studies measured input lags by comparing the difference in play-out between two devices, connected to a similar digital clock input. By recording the two devices the play-out difference was measured by taking the difference between clock times, clearly observed in Figure 2.2. Figure 2.2 shows just by looking that especially the 10 ms indicator is difficult to read out. This is due to the camera device operating at a lower frame rate than the update frequency of this number. This number updates at a frequency of 100Hz while the camera only operates at 50 Hz making the recorded image unambiguous. In the case that a TV-broadcast is recorded a clock signal may not be available. If a clock signal is available this method can also be used to measure the difference between co-located television viewers.

2.2 Analysis of measurement systems

The measurement schemes are compared schematically in Figure 2.3. To develop a full end-to-end play-out difference measurement tool we want to take as many of the differences encountered between two receiver terminals into account. For example if two participants are watching similar content obtained from a different source we still want to know the exact difference in play-out times.

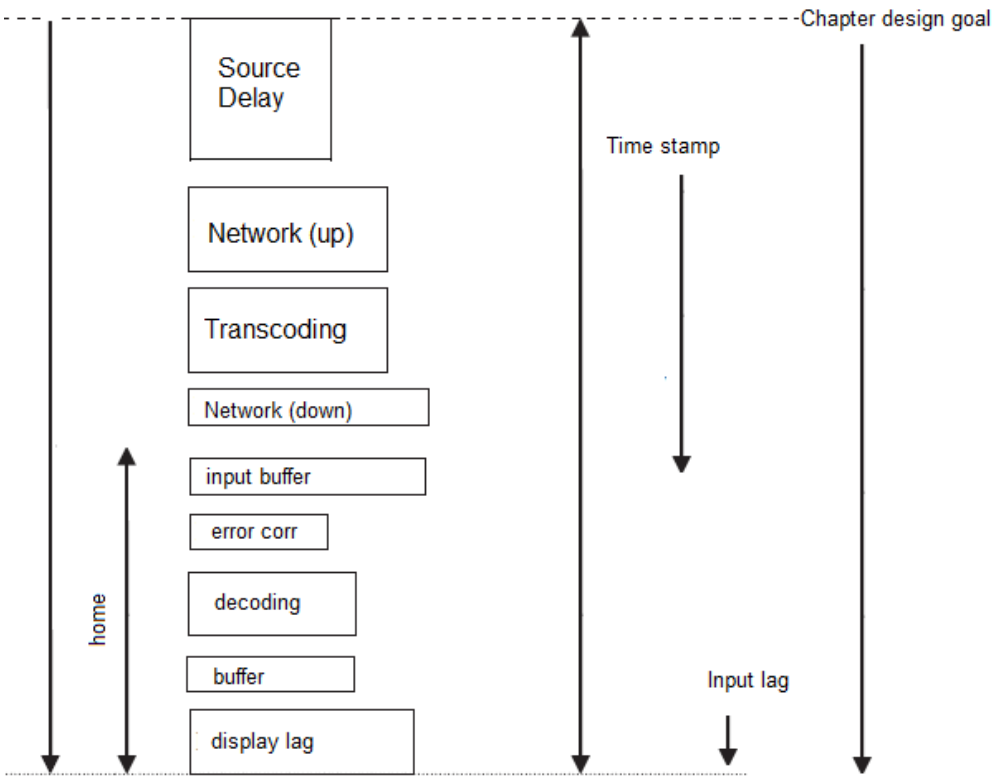


Figure 2.3 End to End delays encountered in multimedia distribution

The time stamp method as in [4] might work well with measuring terminals running similar hardware and similar network/protocol stack. This could be for example in a commercial video conferencing system. These requirements may not hold in TV-broadcasts which have various network types and devices. Also for our research purpose they are not practical as we measure in proprietary networks. Also timestamps are not clearly defined for analogue signals which we do want to measure in our TV broadcast investigation. Another drawback of the timestamp method is that delays introduced after digital reception (timestamp) by both the set-top box and the (digital) TV as shown in Figure 2.4 are not taken into account.

The approach of reading out clocks is accurate but hard to do when measuring actual TV broadcasts as a clock signal is not always available. This makes it more suitable for screen measurements only as shown in Figure 2.3. As the main aim is to measure play-out differences in a pilot study in broadcast TV we will combine the front recording method from section 2.1 to design a new system that fulfills the following requirements:

1. Accurate measurement of play-out difference between two devices
3. Take end delays caused at the home by set-top box and TV into account
3. No access to network or sender
4. Easy to deploy such that many measurements can be taken at different locations

These are requirements for performing a measurement study of play-out differences between TV broadcasts as done in chapter 3. Requirements 1 and 2 are needed to obtain accurate measurements. Requirements 3 and 4 are practical constraints for our specific measurement study in chapter 3. In chapter 3 we will measure at the homes of test participants and proprietary networks. Easy deployment and avoiding access to sender or network will make it easier for people to participate.

These aims are achieved by recording and measuring at the end of the terminal taking delays in the screen and set-top box into account as in [19]. We will trigger on scene changes instead of digital clock times, this way play-out difference can be measured automatically without the need to set the input signal to a digital clock. We will employ automatic scene change detections and use correlation to estimate/measure play-out differences in an accurate and robust way.

2.3 Design of a new measurement tool for inter-destination synchronization

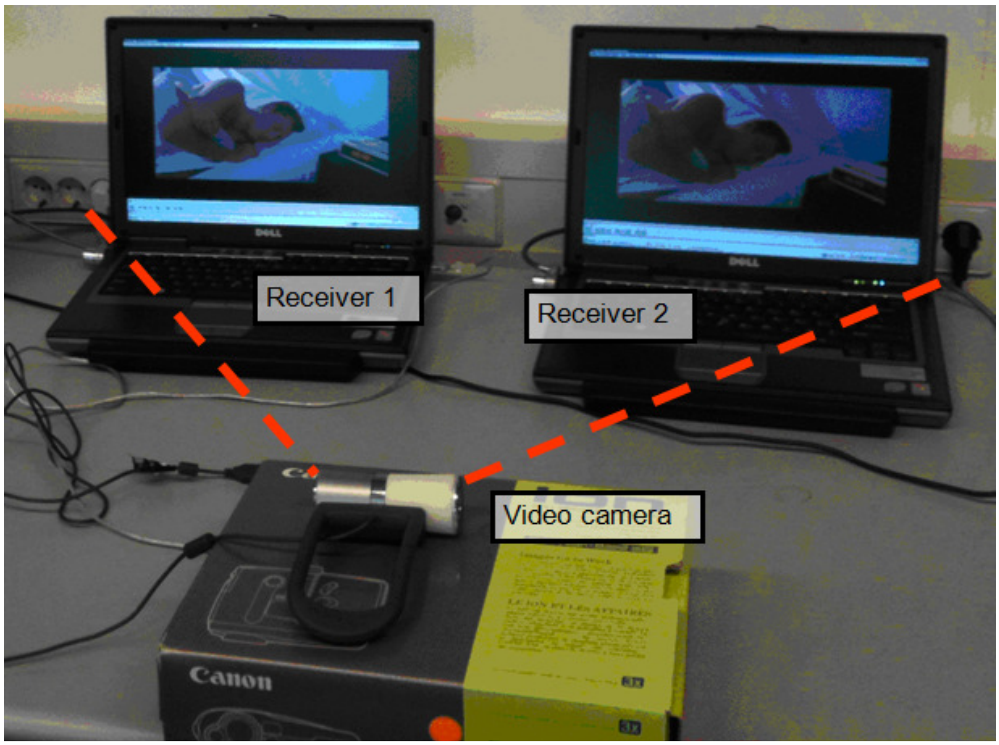


Figure 2.4 Measurement setup for play-out difference measurement

By taking the design considerations into account, we start measuring play-out difference by recording the two subject videos as shown in Figure 2.4. The recording of the two videos is used later to compare them for play-out difference. The camera used for recording the two videos, along with the frame rate of the videos, defines the achievable accuracy of the play-out difference measurement. When a mobile receiver with play-out location invariant play-out difference is used, this method can be extended to indirectly compare non co-located viewers. In accordance to the Nyquist rate, the frame rate of the measurement camera should be at least twice that of the frame rate of the videos to achieve a complete sample. However lower frame rates for the measurement camera may be chosen if a lower level of accuracy is sufficient. For our experiment a Logitech Ultra vision quick cam operating at 30fps was used. At first the approach was to compare scene changes on both screens to compare the play-out times to compute the difference, later this approach was made automatic resulting in the design scheme shown in Figure 2.5.

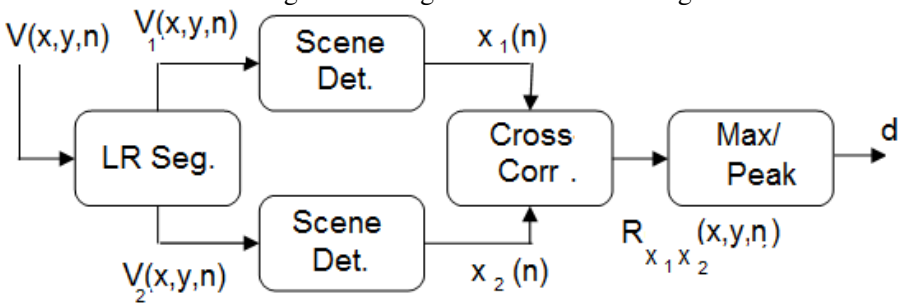


Figure 2.5 Play-out difference measurement system design scheme

In Figure 2.5 V(x,y,n) represents the frame recorded by the camera at the input, the left and right devices are separated in V1(x,y,n) and V2(x,y,n) and subsequently scanned for scene changes.

An estimate of the cross correlation between the detected scenes is used to detect the play-out difference. The estimator for the cross-correlation is an unbiased estimate and given by the equation:

$$R_{x_1x_2}(k) = E[x_1(n)x_2(n-k)] = \frac{1}{M-k}\sum_{n=k}^{M} x_1(n)x_2(n-k) \quad (2.1)$$

Where M is the number of frames recorded and $x_1(n)$ and $x_2(n)$ are the detected scene changes from the two segments and R_{x1x2} is the cross correlation between the detected scene changes and is assumed independent of the time n. The independence on n is assumed as the play-out difference is assumed to be fixed. We assume other sources of cross correlation to be independent of the time n.

A sample cross-correlation from estimated data is shown in Figure 2.6. It shows a clear peak at the play-out difference that is easy to detect. The peak is a little bit wide indicating that occasionally play-out differences changed a little bit.

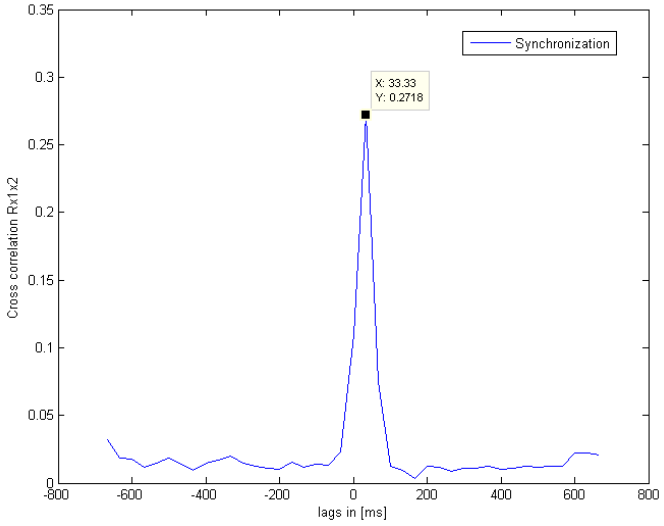


Figure 2.6 An example cross correlation plot with 33 ms play-out difference

The implementation of the scene change detector is given in the next section. The extension to non co-located receivers is given in chapter 3.

2.4 System implementation of [21]

This section presents the MATLAB implementation of play-out difference measurement scheme. We present the algorithm from [21]. To understand how this algorithm works this section first starts with discussing the various image features used in this algorithm in (2.4.1, 2.4.2 and 2.4.3) and their implementations. Then in 2.4.4 the implementation including parameters to tune the algorithm are given.

2.4.1 DC image coefficient difference computation

The DC image value is an average intensity over a block of pixels, 8x8 corresponding to an often used block coding size in compression, equals the first DCT coefficient of 8x8 DCT of that block. The difference in DC image value per block is a good indicator of scene changes [21], these differences can be computed from DC image values by taking the sum of all differences between corresponding blocks.

$$F_{dc}(m,n) = (1/64)\sum_{i=0}^{7}\sum_{j=0}^{7} f(8m+i,8n+j) \quad (2.2)$$

$$Diff_{dc} = \sum_{n}\sum_{m} F_{dc\_X}(n,m) - F_{dc\_Y}(n,m) \quad (2.3)$$

An original image and its DC version are shown in Figure 2.7 with the dc differences computed between frame 1 and frame 2 and frame 3 and frame 2 on the bottom of the image respectively. The DC image basically is a blockier version of its original. The total difference between two frames in DC coefficient is much bigger in case of a scene change as shown in Figure 2.7. The computed differences are shown below in the image.

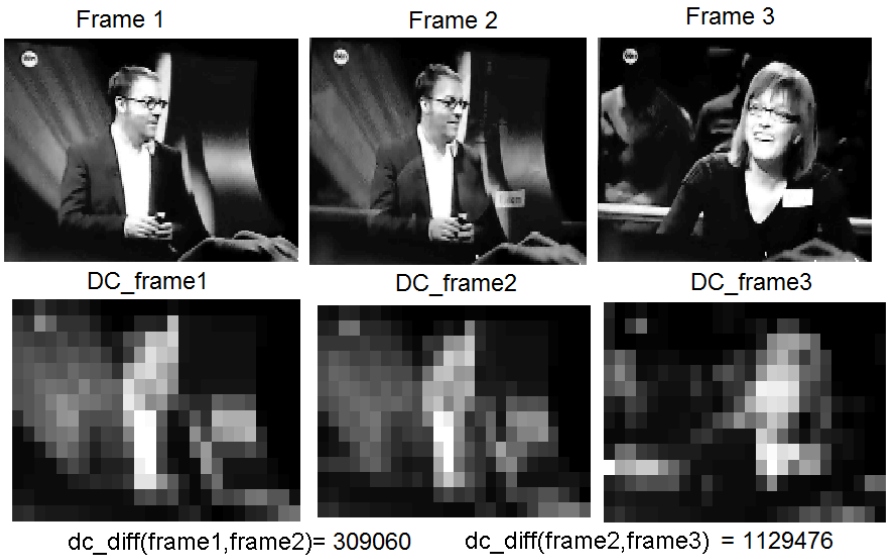


Figure 2.7 Graphical illustration of dc coefficients and differences

The main disadvantage of doing scene-change detection based only on DC image differences is that fast motion yields large differences possibly leading to falsely detected scene-changes (false positives). In Figure 2.8 the values of the DC differences for two segments of video are plotted. The red line represents a segment recorded from the left side and the green line the one from the right side. The peaks in Figure 2.8 mostly correspond to scene-changes but are also occasionally present at high motion fragments. Already from this plot some clear hints of the play-out difference are indicated. The red line represents the video that is ahead. From this picture it can already be seen who is behind and who is ahead, however extension to scene changes made to make the method more generally applicable to different screen sizes etc.

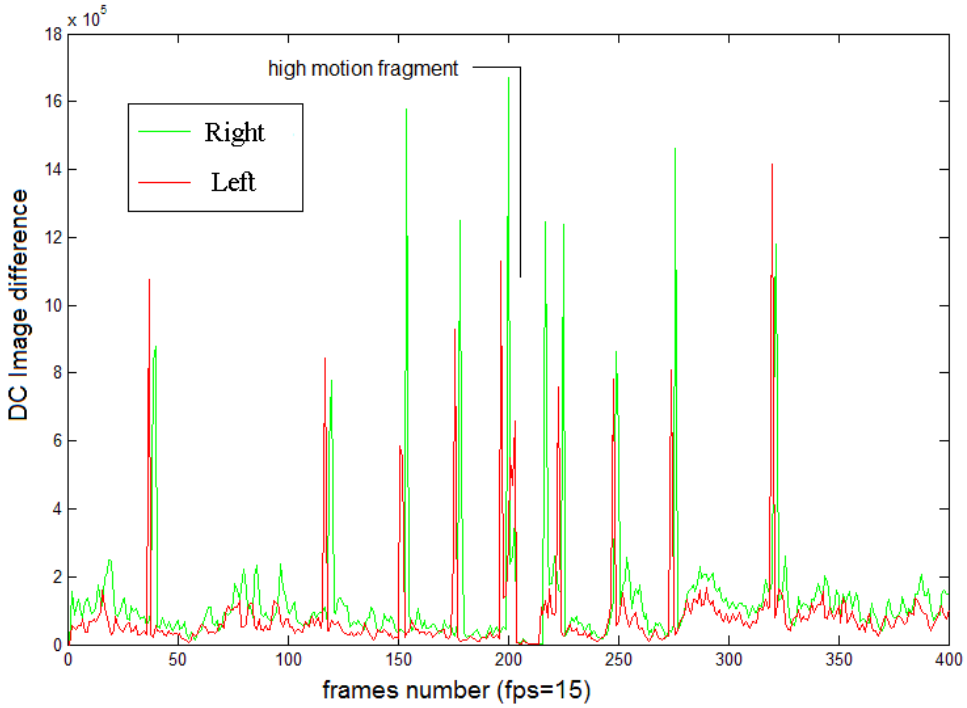


Figure 2.8 DC differences of a sample video

2.4.2 Histogram feature

A histogram of an image represents a range of intensities and a frequency count of how many pixels in the image have an intensity in that specific range. The histogram feature complements the DC image feature in that it is more resistant to fast motion as in this case the approximate distribution of the intensities/colors remains approximately the same. The disadvantage of the histogram is that it is sensitive to lighting changes. The two complement each other well in the scene-change detection problem as lighting and fast motion are the main sources of error when detecting scene changes. The Histogram difference measure is obtained from [21] is given by equation 2.4:

$$d_h = \sum_{i=1}^{M} \frac{H_x(i) - H_y(i)}{\max\left(H_x(i), H_y(y)\right)} \quad if\ H_x(i) \neq 0 \cup H_y(i) \neq 0 \quad and \quad 0\ otherwise \qquad (2.4)$$

In this equation frame x and y are compared, Hx represent the frequency count in bin i of the histogram of frame X. The maximum term below is used for scaling. An example of two frames, a scene change and their histograms obtained from a sample measurement of our data is shown in Figure 2.9. It shows that in the case of a scene change the difference in histograms is quite large.

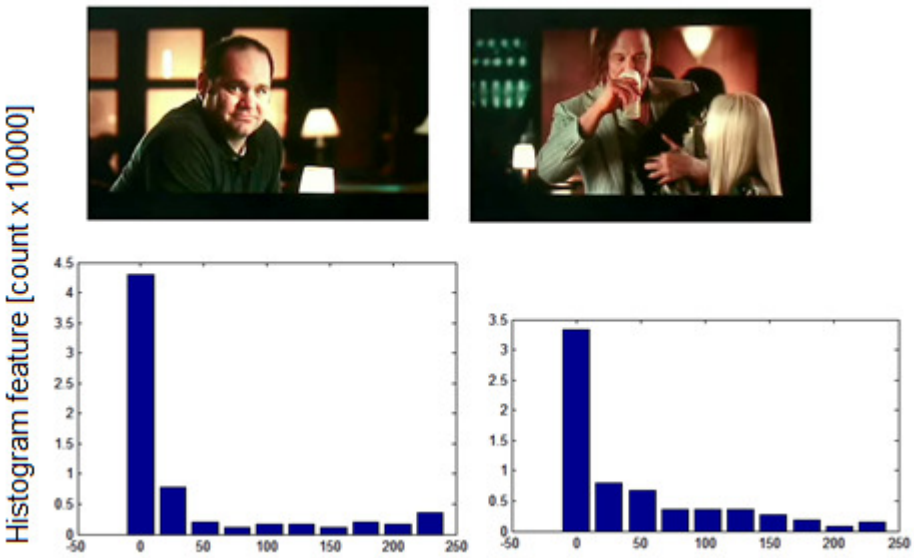


Figure 2.9 A graphical illustration of histogram feature

2.4.3 Static scene change feature

Static scene-change features compare the shape of the objects in the two images. It uses edge detection (detection of borders and lines in the image) to segment the shapes of the objects present in the frames. Assuming that in motion and zooming the shapes of the objects only changes gradually compared to scene changes the application of this feature can be understood. In the computation an edge detector discussed in [22] that is available in MATLAB is used. This edge detector computes lines as a binary image that represent the shapes in the image quite well. For comparing an edge-map of frame X to an edge-map of frame Y, the edge-map of frame Y is made thicker (dilation). Now the pixels of the edge-map of X that fall inside the thickened edge-map of Y are summed up. If the value of this sum is really low it indicates that the objects in the image changed. This implies a value of α approaching 1 indicating a static scene change. The measure is given by the following equation 2.5:

$$\alpha = 1 - \frac{\sum_{x,y} edge(X).dil(edge(Y))}{\sum_{x,y} edge((X))} \tag{2.5}$$

As mentioned earlier the higher α the more likely a scene changes between frame X and frame Y. An example of a static feature detection based on our video data is given in Figure 2.10. The α corresponding to the static scene change feature between frame 1 and 2 is larger than 0.175 (few remaining lines) while the static scene change feature between frame 2 and 3 is larger than 0.175 (many remaining lines)

Figure 2.10 A graphical illustration of the static scene change feature using edge maps

2.4.4 Algorithm calibration

Using implementations of the image features the scene change algorithm from [21] was implemented to detect changes; the algorithm is shown in Figure 2.11. DC-differences between left and right segments are computed using eq. 2.1 and eq. 2.2 in the top of Figure 2.11. After this step, if differences larger than threshold T_l are found, a time-window around this difference is used to compare the difference with the second largest nearby difference in time. By comparing nearby peaks some effects of fast motion are eliminated. If the ratio between the two peaks is larger than the threshold parameter N_r a scene change is assigned. If this ratio is smaller than N_l no scene change is assigned, otherwise the differences are attributed to motion. If the ratio is between the two N_l and N_r parameters the histogram measure is used to decide if a scene-change occurred or not. If the histogram feature is larger than threshold H_R a scene-change is assigned. If the feature is lower than N_l it is rejected. If the histogram feature is between N_l and N_r a static scene change test comparing to T_e is used to either assign a scene-change or decline a scene-change. The algorithm schematic is shown in Figure 2.11

The main challenge was to manually optimize the parameters with the idea of keeping the false positive rate low and the detection probability high. Some movie segments on which scene changes were manually detected were used to tune the parameters. The frame numbers of these manually found scene-changes were compared with the values of the system for various parameters.
We first optimized parameters N_R and N_L were approximately all scene changes are detected, but still a lot of false scene change detections exist. We did this by varying the value until all the scene changes were present in the set resulting from this first step. After this first step a lot of false detections still exist.
After that we tested for the parameters H_l and H_r which removes some of the false detections. After that T_e was tuned to reject the last scene changes were it was not clear whether they were correct or not. The values found are shown in Table 2.1 and are used in each of the measurement conditions.

Parameter	Value
N_R	4
N_L	1.5
T_low	1.5*segment_Area
H_I	2000
H_R	4000
T_e	0.175

Table 2.1 Threshold parameters for scene-detection algorithm

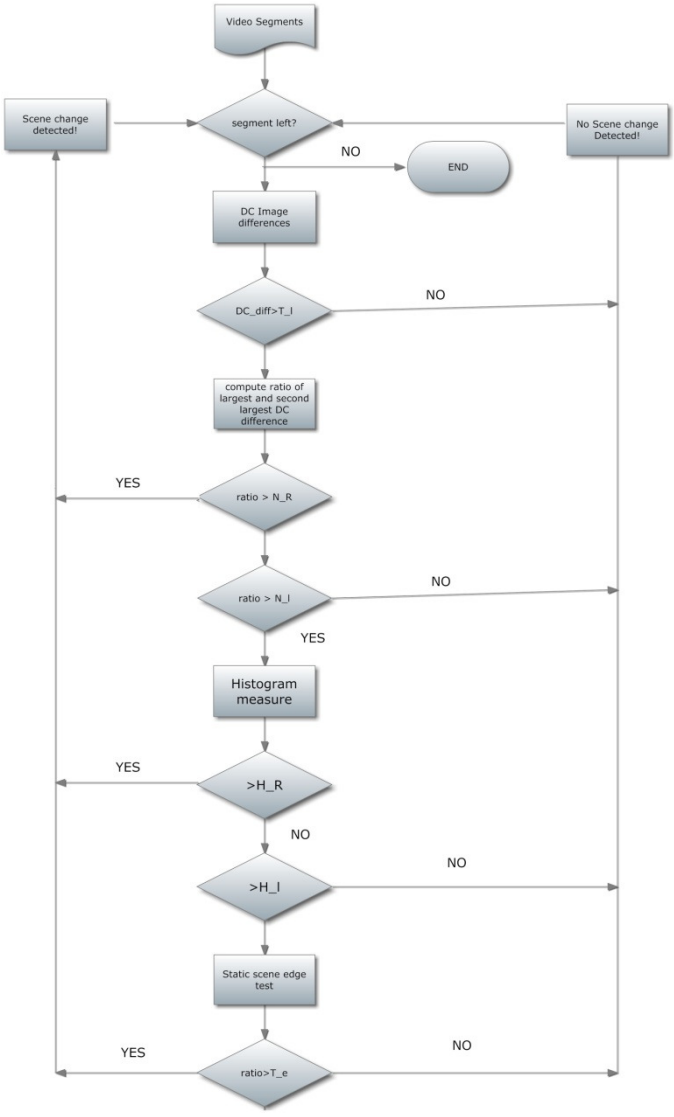


Figure 2.11 Algorithm for scene change detection from [21]

2.4.5 Testing the scene change detector by comparing to a manually labeled sample set

The implementation was tested on a single clip of a heavy action movie “In China they eat dogs”. For 11 different 26 second fragments the results of the automatic detection were compared with a manual check of the scene changes. The count of false detections and misses is given in Figure 2.12 for both the left segments and the right side segments. In both the right and left segments most scene changes were properly detected as indicated by the green area. This indicates proper operation of the scene change detection. The red and gray parts in the graph indicate the errors that are relatively small compared to the correct detections.
We found some false positives mainly caused by explosions and special effects and some missed scene changes. The dataset contained 182 scene-changes (91 on each side) of which 172 were detected. This result estimated an average detection rate of 94.5% and an average false negative rate of 5.5%. The false positives were kept low only 6 out of 8800 frames were falsely classified as scene changes corresponding to 0.14%. These rates are sufficient for our goal of play-out difference detection because we use multiple measurements to detect the play-out difference as will be shown is section 2.6.

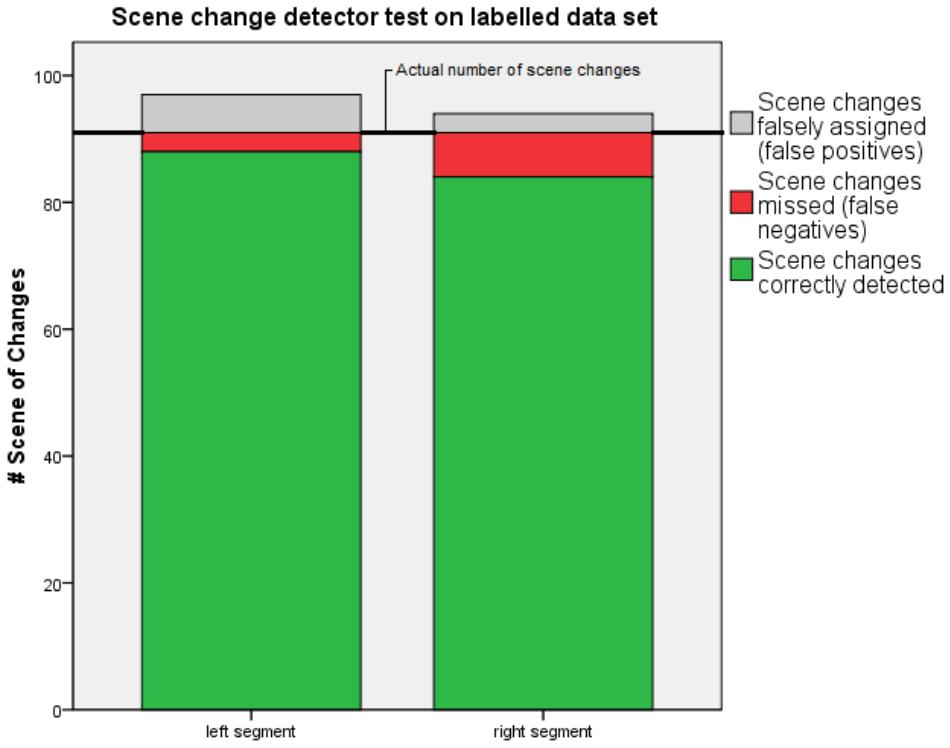


Figure 2.12 System test on labeled dataset showing detection performance

2.4.6 Segmentation and cross correlation detection

The cross-correlation estimate from the detected scene changes was implemented in Matlab using its function xcorr. The left rate segmentation was implemented using the ginput routine from the diplip image processing library [23]. The selection of the left and the right segment is shown in Figure 2.13. The user selects the upper and lower points of the two screen images to segment them from the rest of the picture using the interface with the lines.

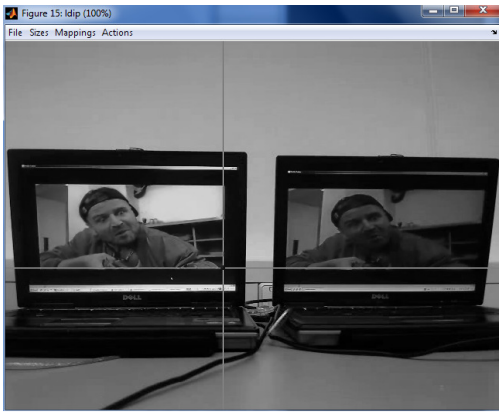


Figure 2.13 Input function for left right segmentation from [23]

2.5 System demonstration

To run the system in practice the file idms_detector_demo is run from the MATLAB environment. The .wmv file with the recording is selected from the dialog box. The user subsequently segments the two different screens in left and right segments to be processed by the scene-change detector. After this the scene changes are computed the synchronization difference is computed form the cross correlation and output on the screen. The demo is graphically illustrated in Appendix E.

2.6 Performance assessment of the new measurement tool for Inter-destination synchronization

We assess the performance of the automatic system by comparing to a datasets of manual measurements. We do this to be approximately sure of the synchronization differences encountered. While both the manual set and the measurements will introduce small measurement errors that are difficult to trace, we consider it as a good first step to validate the proper functioning of the measurement system. A more extensive system test requiring a carefully callibrated synchronization mechanism is needed to verify the system performance more accurately. As we do not have such a system available we refrain from performing this test, also because for our measurement purpose an accuracy within 50 ms is tolerable. The results show that the system is accurate and robust.

In section 2.6.2 we try by evaluating the effects of possible changes in the detection accuracy (caused by lighting conditions and screen sizes) and scene change-frequencies which dependent on the frame-rate and the type of video content what the effect will be on the detection mechanism. We will use a simple mathematical model to illustrate the effect such changes might have.

2.6.1 Comparing the detection system to manual measurements

A pilot dataset of a front recording in 11 different settings with different screen sizes, camera positions and lighting conditions was used for evaluation. 297 manual measurements of frame comparison were performed on this dataset. This way we compare the accuracy of manual measurement to the accuracy of the measurement of the system.
The videos of this dataset were processed by our system. This yielded an average absolute difference of 40 [ms] between the manual and automated differences. The overall average difference between the manual and automated measurements was found to be 6 [ms]. While the system was found to work for different screen sizes and recordings from different angles like the one shown in Figure 2.14, it failed when the recorded image was too dark as seen in Figure 2.14. Another dataset from 3.1 with 25 recordings of two similar devices showed that the measurement consistently gave the same play-out difference measurement.

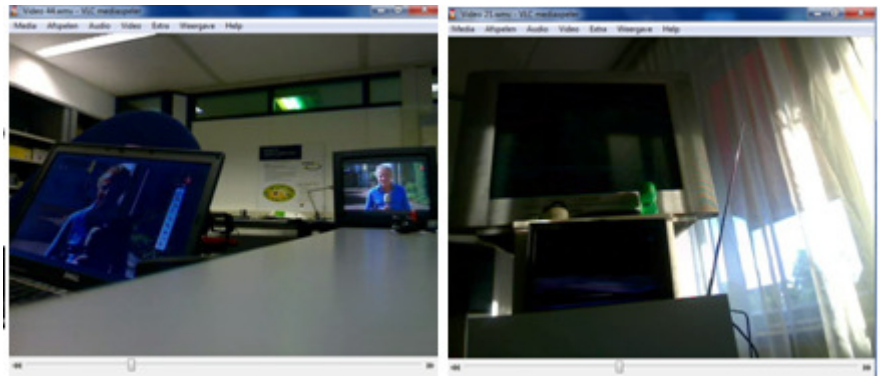


Figure 2.14 Two examples of recordings, a good one on the left and a bad one on the right

2.6.2 Analysis of system performance to parameter changes based on a constant scene-change probability assumption

The algorithm we propose aims to estimate constant or slowly changing play-out differences which we often observed when performing manual pilot measurements of play-out difference between different TV broadcasts. This section evaluates the performance of the system based on p_{f_p} and p_d for the detection algorithm for the case that the scene changes are a realization of P_{sc} (n) with a constant scene change probability P_{sc}. We assume scene changes in a movie occur with approximately constant probability.

While the constant scene change assumption may not be valid in all practical cases, this example shows the effect of using multiple detections for play-out difference estimation/measurement. We expect similar behavior for different behavior for different scene change probabilities, as the analysis will be more complex and beyond the scope of this document and left fot further research.

From the recorded video sequence V(x,y,n) the image of the device on the left side and the device on right side are extracted by separating into smaller videos V_1(x,y,n) and V_2(x,y,n) respectively. We assume that recorded content of devices V_1(x,y,n) and V_2(x,y,n) display similar video-content with an approximately equal play-out rate but a fixed temporal difference as we often observed in trial measurements conducted in the lab. In case of slowly changing play-out difference short samples can be recorded yielding approximately constant play-out difference. Scene changes s_1 in V_1 and s_2 of V_2 are assumed to be versions of the same scene change pattern s(n) shifted by *d* frames:

$$s_1 = s(n), s_2 = s(n-d) \; for \quad n-d > 0 \tag{2.6}$$

The scene change pattern s(n) (1=a scene change 0= no scene change) give as:

$$s(n) = 1 \; first \; frame \; of \; the \; scene$$
$$s(n) = 0 \; otherwise \tag{2.7}$$

As our system is based on the idea that scene changes can be detected with less than 100% detection probability. The detected scene changes in terms of the detection probability and false positive rate are given as:

$$p(x_1(n)) = p_d s_1(n) + p_{f\_p}$$
$$p(x_2(n)) = p_d s_2(n) + p_{f\_p} \tag{2.8}$$

Where the detection probability of a scene change p_d and the false positive p_{f_p} (probability of assigning a scene change incorrectly) are assumed approximately constant and $p_d >> p_{f\_p..}$ From the detected scene changes $x_1(n)$ and $x_2(x)$ of scene changes the cross-correlation is computed which shows a clear peak at the play-out difference if an approximately constant play-out difference exists. The cross correlation between the real scene change pattern defined in eq. 2.9 naturally shows peaks at d as evaluated below using the fixed temporal dependence to:

$$R_{s_1 s_2}(k) = E[s_1(n)s_2(n-k)] = \lim M \to \infty \frac{1}{M-k}\sum_{n=k}^{M} p(s_1(n)=1, s_2(n-k)=1)$$

$$= \lim M \to \infty \frac{1}{M-k}\sum_{n=k}^{M} P_{sc}(n)\, p(s(n-k-d)=1 \mid s(n)=1 \mid k)$$

$$= \begin{matrix} E[s(n)] = P_{sc} & k = -d \\ E[s^2(n)] = P_{sc}{}^2 & k \neq -d \end{matrix} \qquad (2.9)$$

The cross correlation between x1 and x2 shows a similar peak at k as s1 and s2. This can be shown by the following derivation:

$$\begin{aligned} &R_{x_1 x_2}(k) = E[x_1(n)x_2(n-k)] \\ &= E[(p_d s_1(n) + p_{f\_p})(p_d s_2(n-k) + p_{f\_p})] \\ &= p_d^2 E[s_1(n)s_2(n-k)] + E[p_{f\_p}{}^2] + p_d p_{f\_p} E[s_1(n)] \\ &+ p_d p_{f\_p} E[s_2(n-k)] \\ &= p_d^2 R_{s1s2}(k) + p_{f\_p}^2 + 2p_d p_{f\_p} P_{sc} \\ &= \begin{cases} p_d^2 P_{sc} + p_{f\_p}^2 + 2p_d p_{f\_p} P_{sc} & k = -d \\ p_d^2 P_{sc}{}^2 + p_{f\_p}^2 + 2p_d p_{f\_p} P_{sc} & k \neq -d \end{cases} \end{aligned} \qquad (2.10)$$

This implies that play-out difference detection from the correlation function is possible in this case when:

$$P_{sc} >> P_{sc}{}^2 \qquad (2.11)$$

For a constant/uniform scene change distribution with P_{sc} taken into account here this always holds. For good detect ability we determine the bound 10 times:

$$P_{sc} > 10 P_{sc}{}^2 \qquad (2.12)$$

For scene change rates smaller than 1 per 10 frames this should be more than sufficient when constant scene change probability is assumed.
For good detect ability and robustness we also investigate the value of R_{x1x2} introduced at values other than the lag by the terms $p^2_{f\_p}$ and $2p_d p_{f\_p} P_{sc}$ For the detection we would like these values to be smaller than the signal peak $p_d^2(P_{sc} - P_{sc}{}^2)$. For $2p_d p_{f\_p} P_{sc}$ we write:

$$\begin{aligned} &p_d^2(P_{sc} - P_{sc}{}^2) >> 2p_d p_{f\_p} P_{sc} \\ &p_d(1 - P_{sc}) >> 2p_{f\_p} \end{aligned} \qquad (2.13)$$

With $p_d \approx 0.9$ and $p_{f\_p} \approx 0.001$ and $P_{sc} \approx 1/40$ our scene change detector achieved approximately 450 times the minimum bound. For the bound on $p_{f\_p}{}^2$ we write:

$$p_d^2(P_{sc} - P_{sc}^2) >> p_{f\_p}^2$$
$$p_d\sqrt{(P_{sc} - P_{sc}^2)} >> p_{f\_p} \quad (2.14)$$

When neglecting P_{sc}^2 we obtain the performance bound:

$$p_d\sqrt{P_{sc}} >> p_{f\_p} \quad (2.15)$$

With the values $p_d \approx 0.9$ $P_{sc} \approx 1/40$ and $p_{f\_p} \approx 0.001$ the bound from 2.15 is achieved approximately 100 times. This means that even with lower scene-change rates and worse scene-change detector parameters p_d and p_{f_p} play-out difference will still be accurately detected as the peak will still be relatively large. This robustness to parameter changes in the detector is useful as they can occur due to screen-size, lighting conditions or camera resolution. Changes in P_{sc} can also change with the frame rate of the capture device and the type of video content recorded. This robustness is also clearly observed in practice as shown in the previous section.

2.7 Use-cases for the Inter-destination synchronization measurement tool

The system presented can be deployed to estimate input lags of two screens connected to the same input signal. Input lag is still an issue to gamers as shown for example by the activity on the internet forum discussions in [20]. Also not many measurement data is made available by manufacturers, and apart from some amateur measurements [18] consumers have not been able to measure this artifact. The system may need more performance testing to be deployed for official TV input lag measurements.

Testing of inter-destination destination synchronization solutions can be done using the prototype presented in this chapter. Scientific study by [3] analyzing the performance of synchronization algorithm can use the presented prototype, especially in the case of non-homogeneous receivers introducing different delays.

Also recent commercial platforms like clipsync, youtubesocial, Yahoo! Zync can be compared for the synchronization performance by using the presented prototype by third party users. Also the companies can use the prototype for testing their solutions.

In case a mobile TV reference signal is available the system can be deployed to do broadcast measurements between different TV broadcast as will be shown in the next chapter. These measurements are useful for companies providing game interaction (quiz, rating etc.) around TV content as synchronization of the game content to the TV content is needed for fairness. The game content can be phone or internet based. Soccer fans can show interest in the data as it would allow them to choose a TV provider with minimal lag.

2.8 Conclusions and Future work

This chapter presented a prototype that allows simple broadly applicable automated measurements of play-out differences. Compared to methods based on time stamps this method can be deployed with receivers using different video streams, protocols or time references. This makes it very useful for performing measurements for research purposes.

Compared to the input lag measurement, this system does not need a digital clock and is easier to use. The system uses cross correlation estimates to obtain accurate difference measurements. While the processing currently works off-line, a scene change detector in the compressed domain would allow real-time processing. Also more extensive testing and performance analysis of the system on small scale differences is needed to make it applicable for official input lag measurements. In the next chapter a mobile reference video receiver will be tested and the system can be used to compare play-out difference between non-co-located TV receivers.

3 Synchronization differences in Television Broadcasts

3.1 Delay sources causing play-out differences in TV Broadcasts

One of the possible sources of delays introducing play-out difference is trans-coding.
Trans-coding operations occur in the network to adapt to the different capabilities of devices and networks. For example devices with a smaller screen and less bandwidth need a lower resolution and less quality, while high end connections with HD devices require the opposite. To enable transmission to diverse receiver types trans-coding is needed. An overview of various trans-coding techniques is given in [24]. Changing the temporal or spatial resolution, bit-rate, video format, inserted logo or error resilience are all considered as trans-coding operations in [24]. That study also explains that due to computations latencies can be introduced, especially when coding from mpeg-2 to H.264 (reason is that they are based on a different base transform). The internet forum [25] shows how implementers of digital video technologies are struggling to meet latency requirements.

Transmission is another possible source of latency as [6] shows that the DVB-T, DVB-C and DVB-C all use interleaving forward error correcting techniques which have been reported to introduce latency in [26]. The latency is caused because the interleaving procedures are based on rearranging and resending bytes, at the receiver all these shifts need to be received introducing an extra latency.

Delays can also be introduced by buffering in the network or by so called “tape delays” to enable censoring live content. Play-out difference can also exist when different broadcasters simultaneously broadcast content but program their commercials in different ways.

However as these delays can occur on different locations between the sender source and receiver of a broadcast chain we will look at the locations instead of previously mentioned small scale sources of delay. Each location will be seen as a “factor” as shown in Figure 3.1. Delays in the first stage are differences between stations and can be attributed to the broadcast source (different channels). Delays introduced by distribution are caused by the technologies/buffering/trans-coding in the distribution process. The third stage delays are introduced at the receiver end (TV-lag, set-top box lag, home network).

Figure 3.1 Three factors were play-out difference can be introduced

3.2 Measurement Approach

We measure play-out differences by measuring at the homes of users using the tool from chapter 2 and comparing with a reference signal. Our study aims to find the large scale factors that can introduce play-out differences in TV. The 3 large scale factors that we will take into account are shown in Figure 3.1.

The measurements were performed at the homes of consumers of different TV distribution signals such as IPTV, DVB-H, DVB-S, DVB-C SD, DVB-C HD and analogue cable in a geographic region in the Netherlands consisting of the towns: Den Haag, Delft and Zoetermeer. The measurements were performed at 11 different locations in 17 different setups using the front recording method from chapter 2. Measurements were conducted on 3 national broadcast stations in the Netherlands and a local TV station. The data obtained gives a picture of inter-destination synchronization quality between receivers of TV

broadcasts in this geographic area. This is useful for interactive and social TV-services that need inter-destination synchronization. The results are presented in the following sections.

3.3 The DVB-T reference signal

To compare latency of broadcasts at the homes of test participants we used a mobile DVB-T receiver. The DVB-T signal was distributed at 722Mhz (UHF channel 52) on 5 different distribution stations (3 in Den Haag, 1 in Delft and 1 in Zoetermeer) as pointed by a source of the broadcast provider [28]. As they all use the same 720Mhz frequency channel 52 and can interfere with each other we assume that the signals are relatively well synchronized.

Figure 3.2 DVB-T antenna on the roof of a building in delft source: [28]

Figure 3.3 Area under investigation, the locations of the 5 DVB-T antennas are shown [28]

To investigate the effect of signal strength on video play-out latency in the lab a pilot measurement was performed using a digital DVB-T TV with a signal strength measurement indicator. Recordings of the DVB-T TV together with another TV displaying the same channel using an analogue cable signal. 20 measurements of 20 minutes performed with 4 different signal strength levels (obtained by altering or moving/removing the receiver antenna) were provided by TNO. The play-out differences were computed with the system described above and found to be 1.20 seconds 18 times and 1.7 seconds 2 times.

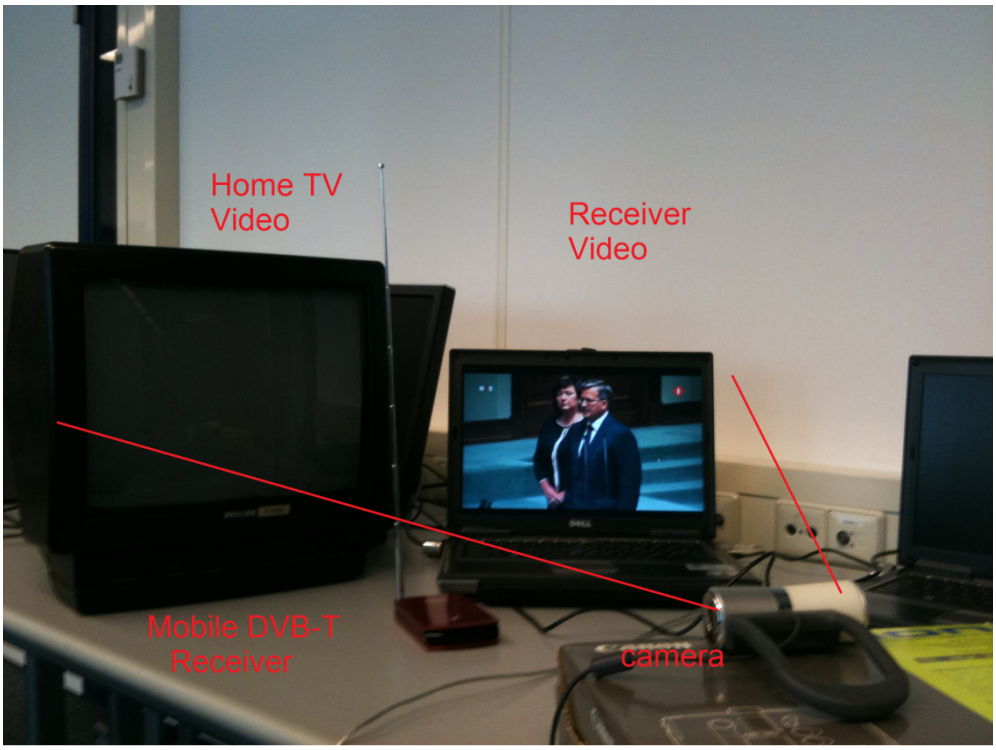


Figure 3.4 The setup with the reference signal, home terminal and camera

As we assume that the 5 broadcast signal sources are relatively well synchronized because they use the same frequency channel and that according to our measurements signal strength does not have an effect on video latency we assume that the reference for this pilot broadcast-TV measurement study is stable and synchronized.

3.4 Broadcast TV

The first study consisted of 297 manual measurements at 11 locations between different TV-broadcasts and the reference DVB-T signal. Virtual Dub [29] was used to compare scene changes on a frame by frame basis. The results of this manual measurement study are shown in Figure 3.5. Figure 3.5 shows the mean difference in play-out between the DVB-T receiver and the technology. Error bars are plotted that also show the standard deviation around the average observed difference. The fact that this standard deviation is low indicates that the difference is approximately fixed and the developed tool from chapter 2 can be applied.

The results of this study also show that play-out difference varies mostly together with the combination of technology (including quality level) and TV station. We tested this statistically using analysis of variance. The test run gave $F(8,249) = 37{,}38$ $p<0.001$ which implies that separating based on technology and channel reduces most of the variance in the measurements. Also per setup approximately constant differences were found, making the use of the automated system valid, especially as some of the remaining variance can be attributed to measurement error. The run of the automated system over the video data is given in Table 3.1 Some small differences were observed, mainly due to the fact of inaccuracy caused by the resolution of the automatic system, or the limited amount of scene-change samples in the case of the manual virtual dub comparison.

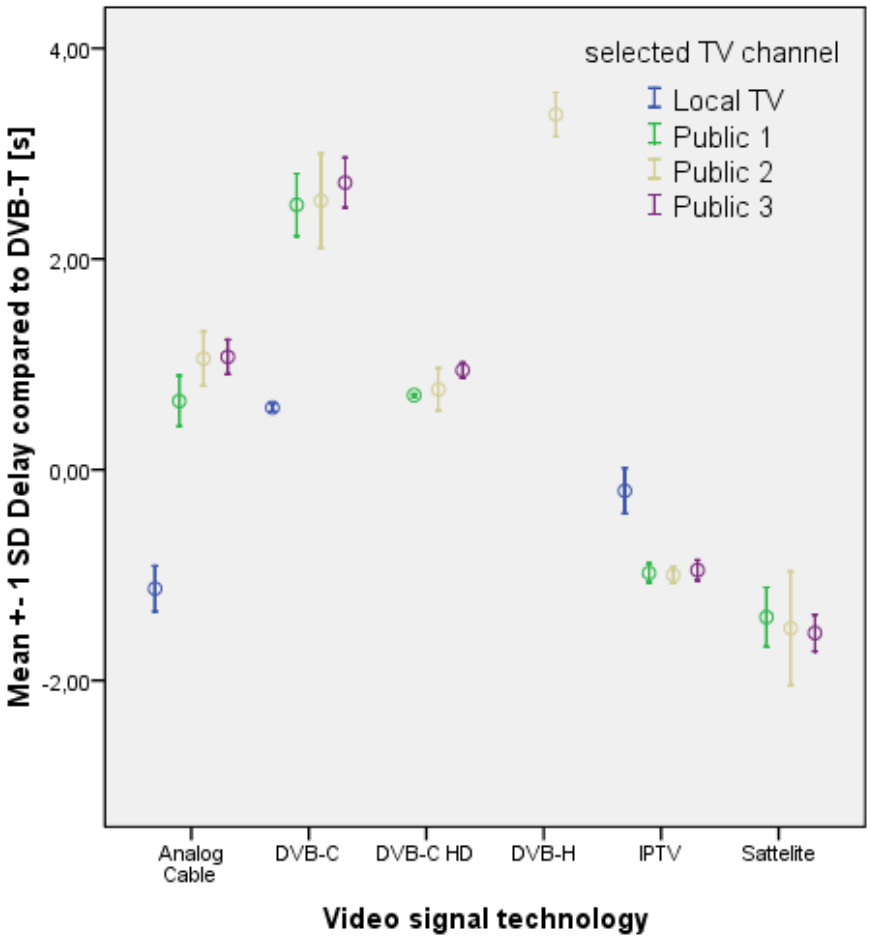


Figure 3.5 The TV results of the inter-destination synchronization pilot measurement study

All tests ran on the same local service providers, except for the IPTV test in which two different companies (KPN and Tele2) existed. For IPTV the effect of background traffic (downloading) while watching TV was tested in one sample measurement. This was found not to have an effect on the play-out time in this particular measurement. Also the difference between the DVB-C SD and DVB-C HD signal is notable. It implies that quality based trans-coding operations can introduce extra video latency.

Manual	Automatic	Technology	Channel	Location
1	0,93	analog	2	Delft
0,93	0,87	analog	3	Delft
0,59	0,6	analog	L	Delft
2,51	2,6	DVB-C	1	Delft
2,68	2,75	DVB-C	2	Delft
2,73	2,75	DVB-C	3	Delft
2,7	2,57	DVB-C	2	Den Haag
0,65	0,6	DVB-C-HD	2	Den Haag
0,52	0,6	analog	1	Leidschendam
1,24	1,17	analog	2	Leidschendam
1,08	1	analog	3	Leidschendam
0,57	0,46	analog	1	Leidschendam
1,01	0,933	analog	2	Leidschendam
1,05	1	analog	3	Leidschendam
1,09	1,2	analog	L	Leidschendam
0,5	0,6	analog	1	Leidschendam
1,21	1	analog	3	Leidschendam
1,18	1,26	analog	L	Leidschendam
0,93	0,87	IPTV	L	Delft
0.9	0,87	IPTV	1	Delft
0,7	0,667	IPTV	2	Delft
0,9	0,8667	IPTV	3	Delft
3,37	3,4	DVB-H	2	Delft
1	1	IPTV	1	Delft
1	1	IPTV	2	Delft
1	1	IPTV	3	Delft
0	0	IPTV	L	Delft
1,4	1,4	DVB-S	1	Delft
1,25	1,25	DVB-S	2	Delft
1,55	1,533	DVB-S	3	Delft

Table 3.1 Field measurements on synchronization difference tested with automated system

3.5 Differences introduced at the receiver

Delays occurring in TV's occur mainly due to enhancement and resolution fitting techniques [30] on digital image input signals. Values of TV lags have been reported to range between 0 and 70 ms [19]. In this study we compare play-out difference between a small CRT-TV (TV 2), a flat screen TV (TV 1) and a large CRT Trinitron TV(TV 3) all connected to the same analogue cable TV signal. We measured the difference for the same three stations with the reference signals as before. As TV 1 (flatscreen) showed the most delay it did not show consistency between the different stations. In this experiment play-out differences were also found too small to be considered a main factor causing play-out difference. The results are shown in table 3.2

	TV 1	TV 2	TV
Channel 1	0.6s	0.46s	0.6s
Channel 2	1.17s	0.933s	X
Channel 3	1s	1s	1s
Channel L	1,2s	1,2s	1,26s

Table 3.2 Comparing play-out difference between TV's

In a second experiment we compared two setups using the same digital cable DVB-C signal but different set-top boxes. Setup 1 uses set-top box 1 and TV 1 while setup 2 uses set-top box 2 and TV 2. We used the system and observed small differences between the two setups. It is not clear that the differences are reproducible and they are small compared to the differences encountered in Table 3.2. Play-out difference cause by the set-top box is therefore not seen as one of the main effects causing play-out difference between TV-broadcasts. However further study using more different set-top box models and setups is needed to jump to conclusions.

	Channel 1	Channel 2	Channel 3
difference	200ms	133ms	66ms

Table 3.3 Comparing play-out difference between set-top boxes

3.6 Web-TV

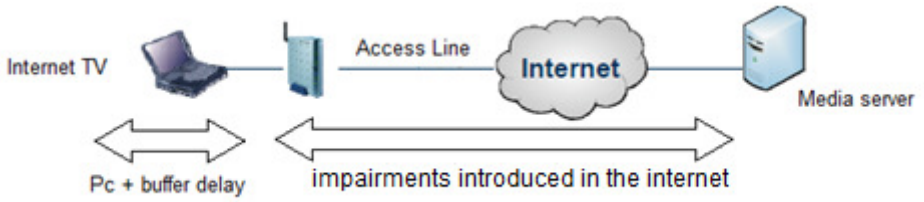


Figure 3.6 Delay impairments in web TV (streaming)

As we saw that in broadcast TV play-out difference seemed to be quite static depending mostly on the technology used and the specific channel. Our next step was to look at web-based TV. As social/interactive applications are more common on the PC-platform than the TV platform, inter-destination media synchronization should be considered for web streams. We measured play-out differences encountered between 3 national broadcast associations in the Netherlands we now look at their website for their web-broadcasts of the national news [31]. At nos.nl news is broadcasted live in a web-browser. The user has the options to watch either a high quality (HD) or a low quality (SD) stream using either media player plug-in or a Silverlight plug-in. In our experiment we connected to the website from two different pc's (pc 1 and pc 2) in the same network and started the broadcast. First we measured differences between the running HD broadcasts measuring 460ms difference between the pc's with pc 1 leading four times. After this we repeated this after restarting the browsers on both pc's. In this case we consistently measured 900 ms difference with now pc 2 being ahead four times. After restarting the browsers again we consistently measured 1.86 seconds different with pc 2 leading. As we figured that synchronization changed when starting the browser we repeated the measurements for combinations with one pc running HD and one SD and both pc's running SD. The measurements were also done for both the Microsoft media player as the Silverlight plug-in. As the values varied quite a lot each time a session was started, they remained more or less constant during a session. Because of the inability to reproduce measurement results after restarting a session only the range of the observed play-out difference observed is tabulated in table 5. The largest differences were observed when comparing a media player HD with a media player SD stream ranging from 4 to 8 seconds.

	HD/HD	*HD/SD*	*SD/SD*
Silverlight	*0-2s*	*2-8s*	*0-2s*
Media player	*2-4s*	*4-8s*	*0-2s*

Table 3.4 Comparing nos web-TV per plug-in and quality level

P2P TV is another example of web-TV that is increasingly becoming popular. Instead of one server distributing video to clients the different terminals also send video to each other. We found one study that assessed the quality aspects such application in [31]. In this study inter-destination synchronization quality was also investigated to vary depending on start up time between 1 and 6 seconds. The large and unpredictable differences between play-out in web-streams show that inter-destination synchronization might be useful here.

3.7 A Network model for IPTV and web-TV

A simplified network diagram representing distribution of IPTV/web-TV is shown in Figure 3.7. Media content is sent to multiple receivers over an IP based network.

In 2007 the ITU presented a standard G.1050 [5] for simulating multimedia network impairments in the network part shown in Figure 3.7. The modeled impairments include packet loss, jitter and one way latency. We will present the results from the standard for latency and jitter which are relevant to our investigation. The model defines three service profiles for different types of IP networks. Levels A and B represent well managed networkes as can be expected in IPTV. Level C presents transmission over an unmanaged network such as the internet.

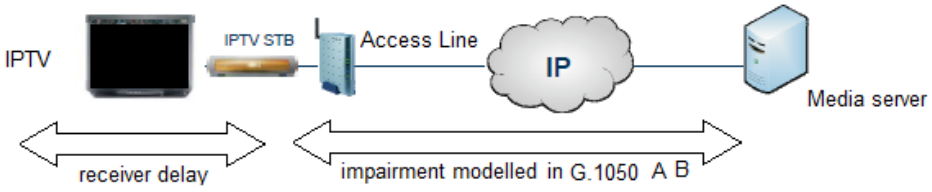


Figure 3.7 Delay introduced in IPTV

Profile A	Regional	Intercontinental
One-way latency	20-100ms	90-300ms
Jitter (peak to peak)	0-50ms	0-50ms
Profile B		
One-way latency	20-100ms	90-400ms
Jitter (peak to peak)	0-150ms	0-150ms
Profile C		
One-way latency	20-500ms	20-500ms
Jitter (peak to peak)	0-500ms	0-500ms

Table 3.5 Jitter and latency impairments encountered in IP based multimedia transmission [5]

The results for latency and jitter are shown in Table 3.5 cumulating jitter and one way latency would yield a worst case inter-destination synchronization quality on the network layer in delivery in the network of approximately 1s in profile C. This value is smaller than the differences seen between TV broadcasts in section 3.2. Also between IPTV services the small play-out difference as observed in section 3.2 can be explained from the data in this standard.

3.8 Conclusions

Figure 3.8 Main factors causing play-out difference

The study performed in a small geographic area showed a range of play-out differences ranging to a maximum of 4/5s between satellite and DVB-H TV signal. The largest differences at the broadcast source were found between the regional and national broadcasts on the cable signals analogue and DVB-C (appr 2s.). At the receiver significantly large differences have only been seen in live web-TV broadcasts from the NOS were buffering seems to make play-out differences unpredictable. The play-out differences between normal TV's and set-top boxes were very small. Most of the differences between video play-out in this study were related to the technology/company used. In the scope of this pilot study play-out differences were approximately fixed if the station and technology were known.

For inter-destination synchronization in broadcast TV the results of this pilot means that simple schemes with an offset per station and channel can be used in broadcast TV and will yield reasonable accuracy. For inter-destination synchronization between web-streams more dynamic synchronization algorithms like those presented in [32] should be employed. The need for inter-destination synchronization within a proprietary broadcast network (The local cable, IPTV etc.) has not been shown to be necessary in this study. For example the G.1050 shows that inter-destination synchronization within a single IP network can be expected to be below 1s which is much smaller than differences between different networks. As the level of inter-destination synchronization required depends on the use case, three use cases will be studied in the next chapters.

3.9 Future work

More measurements are needed to obtain a picture of inter-destination synchronization in a larger geographical context. With the methods and results presented until now a logical step would be to measure synchronization between different DVB-T locations in the Netherlands. From that reference point more measurements can be performed across the country.

This study did not yet investigate differences between stations distributing live content.
A live football match for example can be faster on the British or German TV when compared to Dutch or Belgian TV. These differences happen because of the way broadcasting stations handle coding and transmission of their recorded video differently resulting in different latencies/delays. For sports fans this information can be of interest.

The organization of the TV-broadcasting media landscape can be very different between countries. Therefore the study presented in this chapter should be repeated for different countries/regions were interaction around TV has market potential. The United States is of special interest. The United States has a media landscape with much more large national and small regional TV channels that often display similar national TV productions on different times/time zones and different commercials introducing play-out differences.

4. Inter-destination media Synchronization in Social-TV: a user-test design

This chapter presents a user experiment that measures the effect of IDMS on the user experience in social TV. Proper test design is important to achieve valid results in such a way that they can be analyzed with appropriate statistical methods. If this is achieved it is possible to reject or validate hypotheses formed related to this IDMS effect. This analysis is done in chapters 5 and 6.

The chapter is organized as follows: In section 4.1 we present the basics of quality of experience testing and the first research question. In section 4.2 we analyze the benefits of social TV and present the second research question and hypothesis. In 4.3 the three types of tests are presented from a conceptual point of view: a qualitative, an empirical, and a pilot test for determining test values to use. In 4.4 a use case chosen (genre, episode and device) is selected and analyzed in a systematic way. Section 4.5 discusses the validity of the test and counter measures taken to eliminate threats and guarantee validity are also explained. In 4.6 some other mainly technical implementation details are given. The synchronization algorithm, its validation and the used equipment and software is presented in that section.

Social TV applications and technologies have recently started to provide IDMS using web or IPTV technologies to enhance the shared experience. The deployment of these applications raises questions on their performance requirements. As no previous research on this effect was done, the tests and the results of this test will be new. This is a useful contribution for web-developers implementing web-based social TV solutions. Also broadcast engineers and protocol designers working on interactive TV technology can benefit from the test and the results of the test. On the other hand the design from this chapter can also be used as a starting point for researchers interested in social TV who wish to do a more specific lab based study. Example topics could be for example device (mobile, TV, PC), relationships (friends vs. strangers), or age groups. Also researchers studying how people interact and use media technology can use this test as a base for new research.

4.1 Introduction to Quality of Experience (QoE) defined by the ITU

As we would like results obtained from our user test to be applicable in the telecommunications domain this chapter investigates subjective testing in telecommunications (QoE). Traditionally telecom operators focused on providing good quality of service (QoS) by guaranteeing clearly measurable transmission parameters such as signal to noise ration (SNR), delay, capacity, loss rate etc. However as techniques for transmission and compression changed and became more powerful but more complex, the simple direct relation between the measurable parameters (QoS) and the perceived subjective quality (QoE) was lost. Introduction of new services increased the need for (standardized) quality testing even more. To fulfill these needs the ITU (United Nations standardization organization) consequently published many standards for performing quality measurements. Also in academia research was done on this topic. This section presents some of this work and selects a quality metric for our user-test.

Quality of experience (QoE) is normally used as a term for the subjective quality of a service or application as it is perceived by the end-user, taking all system and network effects into account. The definition given by [33] defines quality of experience as:

Definition 1: Quality of Experience*:"the overall acceptability of an application or service, as perceived subjectively be the end user"*

Note 1 – Quality of experience includes the complete end-to-end system effects (client, terminal, network, services, infrastructures, etc.)

Note 2 – The overall acceptability may be influenced by user expectations and context.

Testing usually involves multiple test-persons in a controlled environment. The most often used measurement scale is the mean opinion score (MOS) which consists of the average of 5 indication points shown in Table 4.1 Judgments can be given as single quality judgment (single stimulus) or a comparison of an impaired version to an original (double stimulus). An example of an important standard for *subjective testing* is ITU-T P.800 for subjective testing of the quality of transmission.

Mean opinion score (MOS)		
MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Perceptible and Slightly annoying
2	Poor	Perceptible and Annoying
1	Bad	Perceptible and very annoying

Table 4.1 Mean Opinion Score

Applying the impairment scale from table 4.1 to inter-destination media synchronization we derive the following research questions on inter-domain media synchronization Quality of Experience.

The first main research question on this topic is if there is a relation between IDMS and this scale, at how many seconds people get annoyed and at how many seconds people get annoyed.

Research question 1a: Is there a relation between IDMS level and mean opinion score? (i.e. do users experience the effect ?)

Research question 1b: at how many seconds do participants start to notice synchronization difference?

Research question 1c: at how many second do participants start to get annoyed by synchronization difference?

These questions are again also relevant for synchronization algorithm designers that are currently mainly using 150ms from [33] as their guideline for seamless synchronization. For the social TV use case tested here, if higher values can be found acceptable then simpler mechanisms can be used instead. For possible adaptation in standards we will use measurement methods similar to QoE testing in telecommunications, as shown in table 4.1.

4.2 QoE of Social TV, the benefits of watching together according to previous research

This section discusses work done in social TV research related to our experiment. It shows that an increased feeling of being together and an improved sense of connection and improved relationships are the main benefits of social TV (4.2.1). In 4.2.2 a case study on conversation patterns in social TV is used to induce hypotheses regarding the effect of IDMS.

4.2.1 Related user-tests, investigating the benefits of Social TV/Watching together

Social TV is currently an ongoing research topic with theoretical research, lab-based and (in home) field-trial based research [34]. The first large scale (in-home) field trial for social TV was ConnecTV [1] conducted in 2007. In [1] 50 households in a city in the eastern part of the Netherlands got a social TV device installed in their homes. The device offered the following services: the possibility to send recommendations, a buddy list, the possibility to switch to a friend's channel and the possibility to switch to the most popular channel. Apart from these services ConnecTV did not offer text or voice chat. In the ConnecTV pilot both a baseline measurement (no ConnecTV functionality) and a pilot measurement (with ConnecTV functionality) were performed. An automatic sampling system was used to obtain information about the user experience directly from the users at home. Because of the automatic system used, this field trial obtained a lot of data on the user experience. The results of the baseline and the pilot measurement were also graphically plotted and are shown in Figure 4.1.

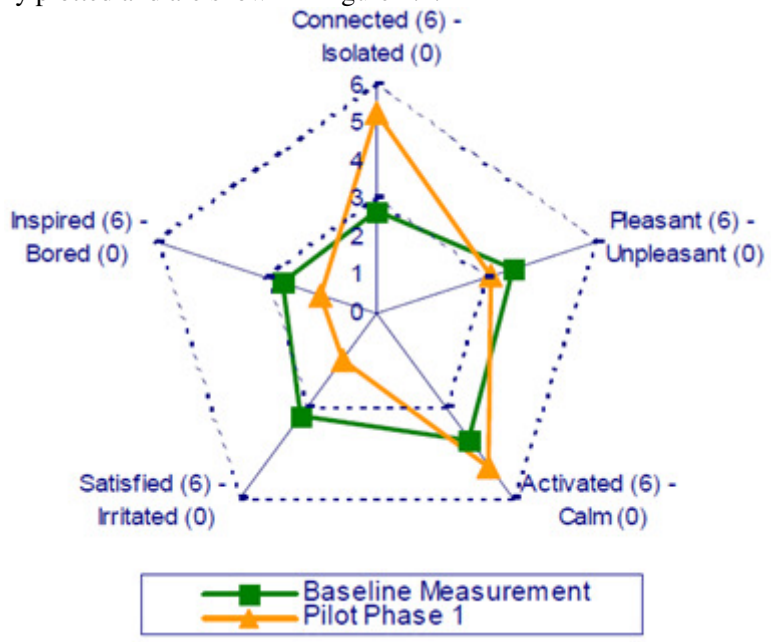


Figure 4.1 User experience of Social TV pilot compared to baseline from [1]

The users felt clearly more connected in the pilot measurement compared to the baseline measurement. As can be seen from Figure 4.1 they also felt more bored and irritated. According to [1] the increased irritation and boredom can be mainly attributed to some technical imperfections and absence of communications functions such as voice or text chat. The desire for these functionalities was also shown in a large scale study on TV systems performed by Geerts [10]. Although it did not offer text or voice chat the ConnecTV pilot clearly showed that social TV even in this case can make people feel much more connected.

In [35] a user experiment on social TV was done to measure the effects of chat on the overall media experience. This experiment was performed in a lab-based environment using laptops. The study compared groups of friends with groups of strangers and both situations with chat and without chat. The overall media experience in this experiment was measured with 4 different constructs on a 1-5 point scale. The

measurement quantities were: joy and fun ("I had fun watching the cartoons), chat enjoyment ("I enjoyed chatting with other people"), closeness ("During the study I felt close to Participant X"), and liking of others ("I liked them"). This interesting experiment showed statistically that the availability of chat or being friends did not have a significant effect on joy and fun. However for both friends and for strangers the availability of chat gave a big significant increase in both liking and closeness (togetherness/connectedness) experienced. The result of the lab based experiment can be seen as similar as in the ConnecTV experiment, the overall experience is not necessarily clearly better but the connectedness is improved.

While [35] considered text chat, a specific lab-based study by Geerts [36] found that voice chat in general is more often preferred than text chat when watching together, the reason was that especially for less skilled users the chat function simply distracts from the video. Some very skilled users of text chat were able to text-chat and watch simultaneously. They actually preferred text chat as they could avoid missing parts of the show because of the other participant's voice interference. These results imply that good lab-based user tests on social TV should in general involve both voice and text chats as they both have an effect on the user-experience.

In Huang et al. [34] an in-house field trial was performed similar to the ConnecTV [15] trial. The Motorola STV3 social TV device which offers both voice and text chat was employed in the homes of 5 externally recruited friends. The test was taken over a period of 3 weeks during a major basketball event. All five participants had indicated to be basketball fans and the researchers hoped to provide a common ground for communications. Compared to the ConnecTV trial this experiment used much less test participants and did no automatic sampling on the user experience. However this experiment collected much relevant information by taking interviews (before during and after the experiment) and analyzing voice and text chat logs. What makes the results of this study interesting is that the obtained information shows how people can integrate social-TV in their lives quite well. Also interesting was that some of the results conflicted with previous results from lab trials. For example the study by Geerts [36] gave voice-chat as preferred medium, in [34] text chat was preferred. Reasons for this were given with phrases like: "*more control of the conversation as no need for immediate response*", "*text chatting takes less energy which I find pleasant after a work day on the phone in sales*", "*not wanting to put someone watching in an awkward situation*" etc. Another result that contradicted previous lab-based studies like [11] was that much of the conversation that took place was not about the content of the show. The recruited participants were friends and had busy schedules in the daytime, they chatted more about the different things in life than just the TV show. The main result obtained from this study that did not contradict previous lab results was the improved connectedness and togetherness experienced by using the social TV device. Many interviews were taken to support this and the one taken one month after the trial when the device was no longer used participants indicated that the social connections afterwards had become much less strong. Some participants therefore really missed the device as they enjoyed the better relation and connection with their friends.

Shamma [34] used a quantitative usage study of three then recently introduced Yahoo! plug-ins providing IDMS to give initial evidence that people feel closer and more connected when watching video synchronized. He presented three plug-ins: Yahoo! Messenger Zync, WebZync and Invisible Zync supporting text chat, voice chat and video conferencing. While the technical contribution here is clear, the argumentation of the effect of IDMS on the shared experience was preliminary and not convincing. Quotes representing the arguments that synchronization improves the user-experience/togetherness are taken of this paper to illustrate that it was inconclusive.

"*In 2007, Weisz, et al. showed that people find the media more enjoyable and feel closer to their peers when they are synchronously watching it with others through Internet chat*"

This is a bit misleading, Weisz [35] only shows that both friends and strangers feel more together when using text chat compared to not using chat. No synchronization conditions were compared (all were watching synchronously), comparing being synchronized to being not synchronized with the same conditions is the correct way to show the effect of synchronization. This type of experiment was neither performed by Weisz nor by Shamma.

"*We began to conduct interviews to explore if people feel more connected with their friends… The interviews help explain patterns found in the log data (such as high chat volumes appearing towards the end of the video and the frequency of emoticon usage during playback)"*

Results of these interviews were not compared to a non-synchronized condition. Also it can be expected that people who download the latest yahoo video plug-ins can be people who are adept chatters and like combining voice/text-chat with video watching, this makes the result only applicable to this specific group. Therefore the recruitment/external validity threat (see section 3.5.2) was not appropriately countered in this article [12]. This makes the results of this paper questionable. The paper also uses testimonials to make a statement that users feel more together
"Let me start by saying, I absolutely love Zync, currently myself and my wife are about 2000 miles apart but we love to watch movies together and it's the closest thing we have to actually being together…. Thanks again we love zync its really made being apart more bearable"
While the testimonial of this user shows the positive effect the tool has on sustaining his/her relation it is narrow evidence for the overall case. The authors of the paper report that they had their tool downloaded by 2,814 unique users. From all these users there will always be some who are very positive. It would have been much better to do a (web based) survey to obtain many measurement data similar to the ConnecTV trial from [1] shown in Figure 4.1. Only by comparing this data with data from a non-synchronized condition appropriate conclusions can be drawn.
Huang et al. also uses quotes, but these quotes came from a small clearly described group and are not used to make overall statements.

Conclusion: *Connectedness/Togtherness is the main merit of social TV According to previous studies. This aspect should be taken into account in a user study on social TV.*

Companies have started offering similar solutions to Yahoo! Zync (proposed by Shamma) for providing IDMS. The most notable are the BBC iplayer [37] for watching BBC programs and youtubesocial.com by socialvisioninc for watching youtube [38] video's together online with a facebook account. Also clipsync.com [39] helps large media companies offer shared synchronized experience. These developments together with previous social TV research highlight the relevance of the following research question:

Research question 2: *What is the effect of play-out difference (IDMS) on the togetherness/connectedness in a social TV application (web-based or home based) measured on a 1-7 scale (Strongly together/connected – Strongly disconnected/ not together) for both text and voice chat?*

Answers of this research will in general be useful for obtaining synchronization requirements for engineers/ web developers working on these social TV applications. For social TV and communications researchers on the other hand it will give more information on the effects of talking around video content. Also companies wishing to start offering social TV services can use answers of this research question as a motivation for developing their business cases.

4.2.2 User behavior in social TV related to IDMS

Interactions in a group watching television were studied by Oehlberg et al. [11] and resulted in design recommendations for social TV. From observations of participants watching in groups, behaviour and conversational structure was analyzed. We believe that the results from this analysis can be used to induce hypotheses about the effect of IDMS/play-out difference on the shared/social experience. We will give a summary of the results obtained in this paper and explain how they relate to IDMS/play-out synchronization.
The study compares the observations of one large group watching TV together to observations of two smaller groups watching together connected by an audio link. The experimental setup of this second case is shown in Figure 4.2. 4 to 2 people in an isolated room in the left were connected through an audio link to a room in the right with 2-4 people. They watch the same programs synchronized while talking with one another.

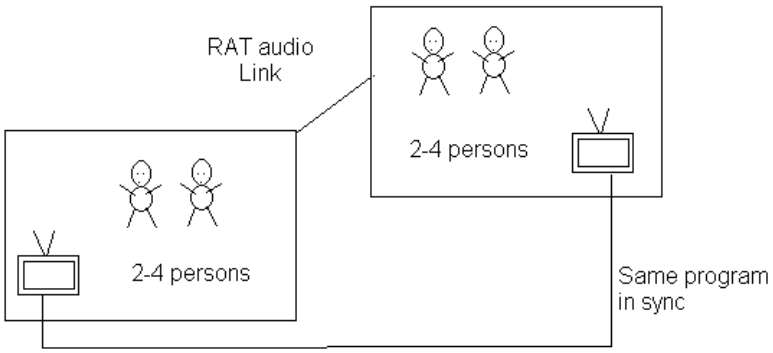


Figure 4.2 Setup from [11] featuring two isolated rooms and an audio link

For both cases (separated in small groups and one large group) some similar observations/conclusions were made on the conversation and behavior we linked to the effects of IDMS. The study gave the following conclusions:

1. *In both cases participants are communicating only at silent pauses, breaks and scene-change as if they are following unwritten rules*

If there are play-out differences in both locations (not the case in this study) this observed mechanism will not work properly. The naturalness of the conversation will be lost, decreasing the quality of the conversation and the experience.

2. *In both cases people talking and listening do not feel distracted from the television show*

If there are play-out differences participants may get bothered and distracted from the television program by comments initiated at silent periods at the other side of the audio link, when the video on the receiver's side is not silent.

3. *In both cases conversation evolved around the content, off- topic remarks were generally seen as awkward by the other participants*

If play-out differences become large, delayed or future content can be seen as off topic and awkward. This should have a negative effect on the overall social experience.

4. *In both cases visual interaction was almost absent. For example during the watching of a football match when a participant dropped onto the ground everybody remained visually focused towards the television*

This could indicate that visual communication is not one of the main factors determining the togetherness/connectedness of the shared experience. From this it could be said that the relative importance of IDMS on the shared experience is relatively larger than expected.

While the observations from this study allow us to induce three hypotheses related to togetherness/connectedness and the QoE (MOS) given below.

The observations from 6 and 7 allow us to formulate hypotheses 1a-1c

***Hypothesis 1a**: A voice chat social TV application resembles watching together at one location quite well when play-out is synchronized (IDMS). Social TV with IDMS and voice will achieve high togetherness/connectedness.*

***Hypothesis 1b:** In Voice-chat social TV synchronization difference will become noticeable and/or annoying because of the described effects. So for the voice case play-out difference have a negative effect on the MOS (QoE)*

***Hypothesis 2**: Play-out synchronization differences have a negative effect on the quality of the conversation. The feeling of connectedness/togetherness will decrease because of off topic comments and worse conversation quality possibly distracting from the show. They will feel less together.*

4.3 User test design

This section deals with the actual design of the user test. This implies choosing explanatory and dependent variables to answer our research question. In section 4.3.1 the experimental design of the test is given together with its conceptual model. As Huang et al. [34] showed that interviews can be very appropriate to obtain information from social TV experiments; we also include a qualitative interview based test in section 4.3.3. A pilot test to select appropriate synchronization conditions is given in section 4.3.3.

4.3.1 Experimental design

This paragraph gives the experimental design for the user test to find the effect of IDMS on the user experience. The conceptual model is shown in Figure 4.3 with independent/explanatory variables on the left, a block in the lower middle illustrating interaction effects, a block in the upper middle showing controlled variables and the blocks in the right representing the user experience.

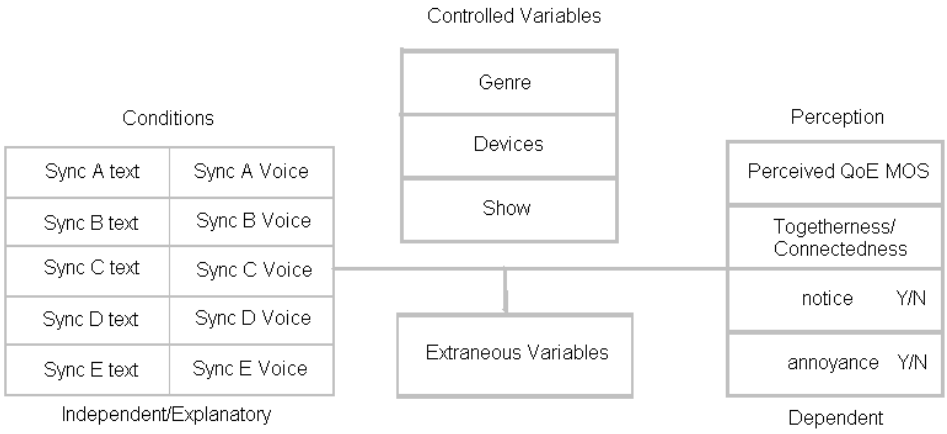


Figure 4.3 Conceptual model of experimental design

The use of text or voice is chosen as one of the explanatory variables as it was previously shown to have a large effect on the user experience [36]. Difference between participants is another effect that was reported to be large in social TV systems. This model takes a within subject approach which implies testing each condition on each participant. This reduces possible bias introduced by differences amongst participants. Repeated measures analysis (ANOVA, Cochrane's Q) or paired samples testing (two samples t-test, McNemar) will be used to explicitly account for the variance between participants in later statistical analysis. 5 IDMS synchronization (play-out difference) conditions are chosen as explanatory variable, they are needed to validate or reject hypotheses concerning research questions 1 and 2. The actual synchronization values to be tested will be derived from a pilot test which is a test taking place before the actual user test. This test is described in section 3.4.2
In the top middle a box with controlled variables is shown. These variables can have a large effect on the dependent variables and are therefore kept the same at a predefined value derived in 3.4.
The box in the bottom middle represent extraneous variables, they can be chat experience, age, sex or for example education level. These variables can have an effect on the relation between the independent and dependent variables and are not controlled for (i.e. kept the same). Later analysis will investigate these effects which are all measured in a questionnaire at the end of each user test. The effects of these variables are assumed less than the controlled and explanatory variables; this will also be tested statistically later. As for the dependent variables shown in the right, MOS score is measured with an adapted DCR (degradation category scale) as shown in table 4.1. These measurements will be used to evaluate the first research

question and will provide results consistent with other perception experiments and standards in telecommunications. The binary variables “the difference was noticeable” and “the difference was annoying” are derived from the MOS indication and are used to find approximate thresholds for perception and annoyance.

4.3.2 Pilot testing

For determining our test-values of interest for IDMS a pilot test is performed, first with members from the research group. This pilot test will give us value for the possible detection threshold for IDMS. From this value we will take 3 values around it to try and find the threshold of perception. 0 and a large value at which the difference is almost certainly noticed are also used in our main experiment to investigate the QoE at the extreme values.

The following pictures indicate how the pilot test is performed. The administrator controls the synchronization difference while the test users **pilot user 1** and **pilot user 2** do not know the synchronization difference (Figure 4.4).

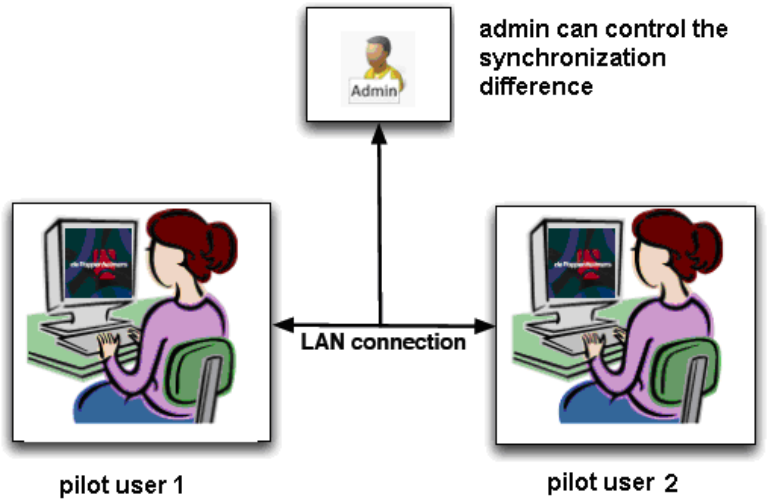


Figure 4.4 Pilot testers performing test to find appropriate test values for the main test

Pilot users 1 and 2 press j at concrete events in the show (questions and answers in the displayed quiz) and try to see if they are synchronized and how much the difference is. The administrator will adjust the values until the participants are just able to detect the difference (just noticeable difference). This value is used as the outcome of the pilot test as approximation of research question 1b. A typical chat sequence obtained from an actual pilot test is shown in (Figure 4.5). If users feel they are not synchronized it can indicate the difference detected. The detected difference was found to be between 500ms and 1000ms (800ms). We chose 3 values around this value 500ms, 1000ms and 2000ms to use and two more extreme values 0ms and 4000ms to be used in the final experiment.

<17:54:07> Chloe: j
<17:54:09> bob: j
<17:54:12> bob: j
<17:54:15> Chloe: j
<17:54:20> bob: j
<17:54:31> Chloe: j
<17:54:36> Chloe: j
<17:54:36> bob: j
<17:54:41> Chloe: j
<17:54:41> bob: j
<17:54:49> Chloe: j
<17:57:08> Chloe: j
<17:57:10> bob: j
<17:57:16> Chloe: j
<17:57:18> bob: j
<17:57:21> Chloe: j
<17:57:24> bob: j
<17:57:27> Chloe: j
<17:57:30> bob: j
<17:57:33> Chloe: j
<17:57:42> Chloe: j
<17:57:44> bob: j
<17:57:49> Chloe: ok, twee seconden denk ik
<17:57:54> bob: Denk ik ook

Figure 4.5 Chat log of pilot user trying to detect synchronization difference by pressing j

4.3.3 Qualitative research (flexible design)

To get knowledge on why users perceive the social TV application as they do, users are interviewed. The questions can be used to induce hypothesis for our previously formulated research questions. Each question is marked with the research question they relate too. Question 3 is used to check the differences between voice and text chat as discussed in [36], allowing us to indentify good chatters. The questions are asked at the end of each test session. The advantage qualitative research is that underlying reasons can be identified. Qualitative research does not measure MOS or any other data that can be statistically analyzed. The main motivation is to research the underlying reasons for way users perceive the application and interpret their behavior. In this thesis the results of this questionnaire are only used.

Interview questions:
1. How did you feel compared to watching together on the couch together?
2. Did you notice synchronization differences and how much did they disturb you?
3. What did you like more voice or text chat and why?
4. Would you like to do this more often for example if your friend/partner was abroad?
5. Where there other things that bothered you when watching the show?

4.4 Use Case Selection

The use case chosen in the user test was a popular quiz called the Pappenheimers. This use case was chosen by dr. David Geerts. This section analyzes this decision by comparing it to other possible choices.

We aim to perform our test using a fixed genre, show and device to limit the available resources (time, amount of participants etc.). To account for large variance between participants in social TV as observed in [23] each participant is tested in each condition. This makes it possible to use related statistical methods that explicitly take the within and between variance in account. In Figure 4.3 it can be seen that already 10 conditions have to be tested on each participant. If we decide to also test different genres and devices the number of conditions for each person would be at least (10x2x2)= 40. As these are too many conditions for one person (one condition takes approximately 10 minutes) the research is narrowed down to a single genre, device and show. Genre, device and show are chosen in a way which is thought to best represent the broad spectrum of social-TV use cases. The second main criterion for selection is that social interaction should be enhanced as we still believe that only people involved in an active and real conversation will be affected by play-out differences. A third criterion is that a certain time aspect (progress in time) must be present in the show and genre.

In 4.4.1 we define the criteria used for selecting the appropriate genre and perform the selection based on these criteria, in 4.4.2 the specific show is chosen in a similar way. Finally in 4.4.3 the device is chosen from three options: mobile, PC or TV.

4.4.1 Choosing a Genre

Already in 3.2 it was shown that for games requiring quick reactions differences of more than 100ms can be critical and perceived as unfair. The use case of the experiment proposed in this thesis however is slightly different; it aims to find the effect of play-out difference in a relatively normal social TV conversation evolving around the content of the show. This experiment is not about fairness in games or reaction time but on the overall effect of IDMS on the overall social experience perceived in Social TV.
To choose an appropriate genre we formulated 4 criteria for selection. The first criterion is that the genre encourages conversation (social enhancement). Genres during which people talk a lot are therefore preferred. The second criterion posed is that the variance of the amount of content in the time should not be too large. Similar amounts of content are desired in each test interval of approximately 10 minutes. A negative example could be for example a football match which can contain long periods with little action and also periods with lots of action. A third important criterion is that we prefer to use a genre which everybody likes approximately as much compared to a genre which some people like very much and some people don't like at all. The reason is that we want to reduce the effects of likeability on the user experience as much as possible. The time factor is a fourth and important practical consideration taken into account and refers to how content ages over time. Yesterday's news is old news and not suited to be watched today. Converting and recording new episodes during the one week experiment is undesired from a practical point of view and also introduces an extra bias in the form of differences between episodes.
The final analysis and grading is shown in Table 4.2. For evaluating the amount of talk and likeability the results of a related study by Geerts et al. [40] are used. For evaluating the content variance and time factor common sense and reasoning is used.

	Amount of talk	Content	Likeability	Time factor	average
Soap	5	4	2	5	4
News	5	3	5	1	3.5
Quiz	4	4	5	5	4.45
Sports	4	2	3	3	3
Reality show	3	2	3	5	3.25
Talk Show	3	2	3	5	3.25

Table 4.2 Comparing genres, 1 indicates high variation, 5 low variation based

4.4.2 Choosing a specific show to use in the test

An important question that remains is: what is an appropriate quiz show to use in our user test? We will first pose the criteria that make specific show appropriate

1. Most people know it and like it. This criterion is used to reduce variation between participants in overall enjoyment.
2. Content offered by program is time sensitive (to make the IDMS aspect relevant)
3. The show stimulates relatively normal social interaction
4. Content is relatively equally spread, to avoid variation between time slots in the user test
5. The program is suitable for our entire broad participant audience (similar as 1)

The first show proposed as use case is: *Who wants to be a millionaire?* This is a worldwide known game show in which a participant has to answer questions and can win up to a million dollars in the end. We propose this show because it is well known, the content is relatively constant and the timing aspect is clearly present. A second use case possible, were content and time-aspect were clearly present was the Dutch TV show called Lingo. This show has been broadcasted for over 10 years and is well known to the Dutch speaking audience for our user-test. It involves guessing words which stimulates interaction between participants. A third option is the quiz show called the Pappenheimers. The Pappenheimers is a very popular quiz show in Flanders which combines many social, cultural and humorous aspects in a quiz show format. To make a final decision we have to choose the best one of these three appealing quiz-shows. As they all seem well suited for our purpose we have rated the applicability to each of the criteria in table 4.3 by investigating the characteristics of the program. We need to have one good use case as we really want to study the effect of IDMS. We cannot choose multiple shows and or genre as the time and resources for the experiment are limited. We want to find the effect of IDMS in a relevant use case: a likeable show that supports conversation.

	Like/know	Time sensitive	Rel. Social	Equal spread
Millionaire	5	5	4	5
Lingo	5	5	4	5
Pappenheimers	5	5	5	5

Table 4.3 Candidate shows for the test

From the table we can see that the Pappenheimers scores slightly better on the last three criteria. We will shortly explain the reasoning behind this. In the Millionaire and Lingo show answering the questions or solving the correct word are the main objectives. As in the Pappenheimers questions often involve humor and personal preferences of the participants, viewers are more likely to have other normal conversations. For criterion of equal content spread, all shows are appropriate. Also content variance can be improved by pre-editing the show. For the applicability to the broad audience, the Pappenheimers seems a better option than the other two shows because of the presence of humor and cultural aspects making it more appropriate for our broad audience. The popularity of the Pappenheimers in Belgium is also one of the main arguments

of it being suitable to a broad (Flemish) audience. The high likeability was later confirmed by results from our post-test questionnaires.

4.4.3 Choosing the device

For choosing the device there were 3 options: mobile, TV and PC (laptop). Laptop was chosen as it seems to give a user experience with similarities to both the mobile phone (interactivity) and the TV (large screen) as it has a big screen like a TV and interactivity like a mobile smart phone. Further research should look at different genres

4.4.4 Conclusions and future work

A laptop based experiment with the Pappenheimers quiz for a Flemish audience is proposed. This use-case scores high on all our criteria: likeability, social interaction, constant content and also it is appropriate for a broad audience. The results of this experiment should give good general recommendations on IDMS and our research questions. Future research should focus on different genres and devices as they have previously been shown to have an effect on the user-experience.

4.5 Validity considerations

This paragraph considers the validity of our proposed experimental setup. Threats to internal, external and construct validity as explained in [41] are systematically checked against our experimental setup. Threats that are identified to be large enough to have an effect on the overall validity are presented here together with counter measures taken to eliminate them. In section 4.6.1 internal validity is taken into account, external and construct validity are discussed in 4.6.2 and 4.6.3 respectively.

Threats to Experimental Validity (Internal validity)

Internal validity in general refers to things that change in the participant environment during the test other than the independent variable that have an effect on the dependent variable. Examples could be an effect of a pre-test or instrumentation used during the experiment. Another well known effect in tests with repeated measures is the expectation/habituation or regression effects where the order of the applied conditions and the time the participant is in the test has an effect on the participants answers regarding the dependent variable. Identified threats are shown in table 4.4

Threat	Description	Solution
Fatigue/Regression	Answers given later are less extreme	randomized ordering
Diffusion	Participants influencing each others decision	agreement with test user
Mortality	participants drop out of the experiment	agreement with test user

Table 4.4 Identified threats to Internal Validity

The fatigue effect is the situation that a user, after being exposed longer will start to give less extreme answers. The regression effect is about differences between conditions, for example in an audio test hearing a soft sound is harder after hearing a loud sound than the other way around due to habituation effects. We expect similar effect for the synchronization conditions. The countermeasures taken against this effect is to vary (randomize) the order of the synchronization conditions and the order of text chat and voice chat. Table 4 shows the randomized time order schedule for our experiment.

Pair	sync1[ms]	sync2[ms]	sync3[ms]	sync4[ms]	sync5[ms]
1	+500	-4000	-1000	0	+2000
2	-2000	-4000	+500	+1000	0
3	-4000	+2000	-500	0	+1000
4	+4000	0	+500	-1000	-2000
5	0	-500	+2000	-4000	+1000
6	+2000	+1000	0	-4000	-500
7	-4000	-500	+2000	+1000	0
8	+1000	0	+500	-2000	-4000
9	+4000	-2000	0	+500	-1000
10	-1000	0	-2000	+500	+4000
11	+1000	-500	+2000	-4000	0
12	-1000	+2000	+4000	-500	0
13	+4000	0	+1000	-2000	-500
14	+4000	0	-1000	-2000	+500
15	+500	0	-1000	-2000	+4000
16	+1000	-4000	+500	-2000	0
17	-4000	-2000	+500	0	+1000
18	-4000	0	+1000	-500	+2000
19	-1000	-500	+2000	0	+4000

Table 4.5 Randomized sync conditions

The diffusion effect is error caused by users talking to each other or complaining to each other about the synchronization difference. In this way if one user detects the synchronization difference it is likely that the other will also come to know it. The users are briefed before the experiment not to talk about the synchronization issue and to give their own opinion in the survey. With this precaution taken it is assumed that diffusion effect is eliminated.
The last identified threat considered participants dropping out (mortality). In the briefing participants are carefully instructed and sign a contract for participation. In return participants receive a voucher with a value of 30 euros. With these measures the mortality threat is assumed to be eliminated (we assume the participants keep their promise and keep participating).

Recruitment Validity (external validity)

The main threat to external validity was identified to be related to the recruitment of test participants. As we believe social TV, like normal TV suits a very broad audience we would like our test panel to represent this as well. However time and budget constraints often make it difficult to let a social TV test panel represent a broad audience. For example Huang et al. [34] used 5 men in their thirties with a busy schedule. Another example is the lab based study by Weisz [35] involving mainly young college students with an average age of 24.3. Even large scale field trials like ConnecTV, measured in 50 households were not representative as mainly highly educated men in the age range of 20-35 took part. The study by [12] was another one that involved many participants, however as these participants all took the initiative to download and use the latest yahoo plug-ins by themselves they can be seen as a rather selective group of instant messengers and therefore they do not really represent a broad audience. In this experiment we really want to find reliable thresholds for getting a good QoE of IDMS representing a broad audience. Therefore recruiting a broad audience is one of the main priorities. A second important requirement/wish is enhanced social interaction. We believe that only people involved in real and active conversation can notice or get annoyed by the play-out differences. To achieve this it is decided to recruit couples as recommended in a study by Shrimpton-Smith et al. [42] to enhance the sociability. An external agency was hired to recruit a

broad audience for our test. Figure 4.5 shows the spread in ages, gender and relationships. This spread makes this experiment one of the broadest lab-based social TV experiments so far based on 72 participants.

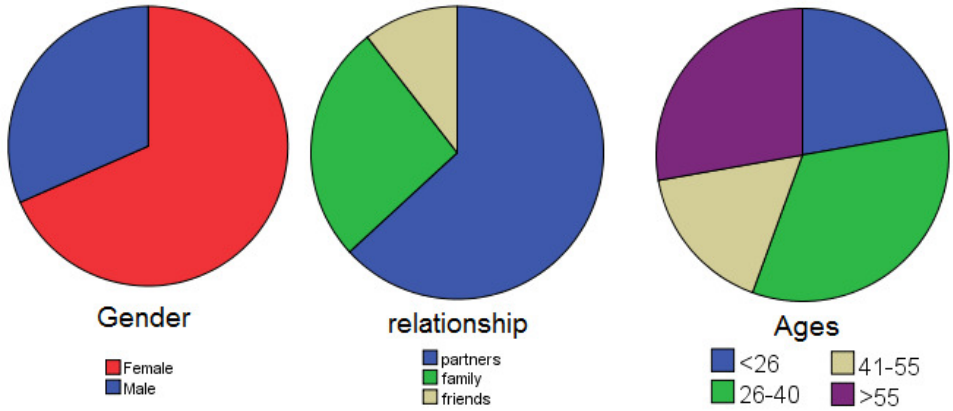


Figure 4.6 Spread of the recruited participants based on age, relationship and gender

Measurement validity (Construct validity)

The Construct validity threat consists of measuring wrong constructs. Because in practice it happens that mistakes are made, attention is paid to this aspect.
We will use the impairment MOS scale to work on research questions, as these scales involve perception, annoyance and MOS scores. We will use the DCR MOS scale defined in ITU P.800. After each test period the user puts in values that will be used in later analysis.

MOS Value	Impairment
5	the synchronicity difference is not perceptible
4	the synchronicity difference is perceptible but not annoying
3	the synchronicity difference is perceptible and slightly annoying
2	the synchronicity perceptible and annoying
1	the synchronicity difference is perceptible and very annoying

Table 4.6 MOS for QoE measurements

Togetherness/Connectedness will be measured by using 6 questions on how connected/together the participants feel. 3 of them are positively posed and 3 of them are negatively posed to cancel for indifferent participants that do not participate properly. If these questions are consistently answered, which can be calculated with reliability indices such as Gutmann's split half and Cronbach's alpha we can assume we validly measured the same construct. The questions are given below and can be answered on a 1-7 strongly agree to disagree scale.

1. The contact with my partner was superficial
2. I felt together with my conversational partner
3. I felt that my partner did not understand me well
4. I felt connected with my conversational partner
5. I got little enjoyment from the conversation with my partner
6. I felt that my partner and I could talk well together

4.6 Implementation details

This section deals with implementation details of how the experimental test was performed. The algorithm enabling play-out differences is given 4.6.1 while the locations and the used equipment is shown in 4.6.2 and 4.6.3 respectively.

4.6.1 Synchronization algorithm Validation

For the experiment a synchronization algorithm implemented by dr. Ishan Vaishnavi from the Amsterdam center of mathematics and computer science was applied in a local LAN. This algorithm made it possible to control the synchronization difference between two users controlled by an administrator (similar to Figure 4.4). One single network time (t) was used from a local web server in the LAN. The local-lag algorithm was published in 2004 [43]. The working is illustrated in Figure 4.7. The master sends its time plus a delay estimate β together with its media play-out position (m0) plus a small advancement to compensate for the delay δ. After user 2 receives the tuple <t+ β,m+δ> it updates its play-out state to m+ δ at t+ β. Note that in our application tn+1=tn+30s indicating synchronization updates each 30s.

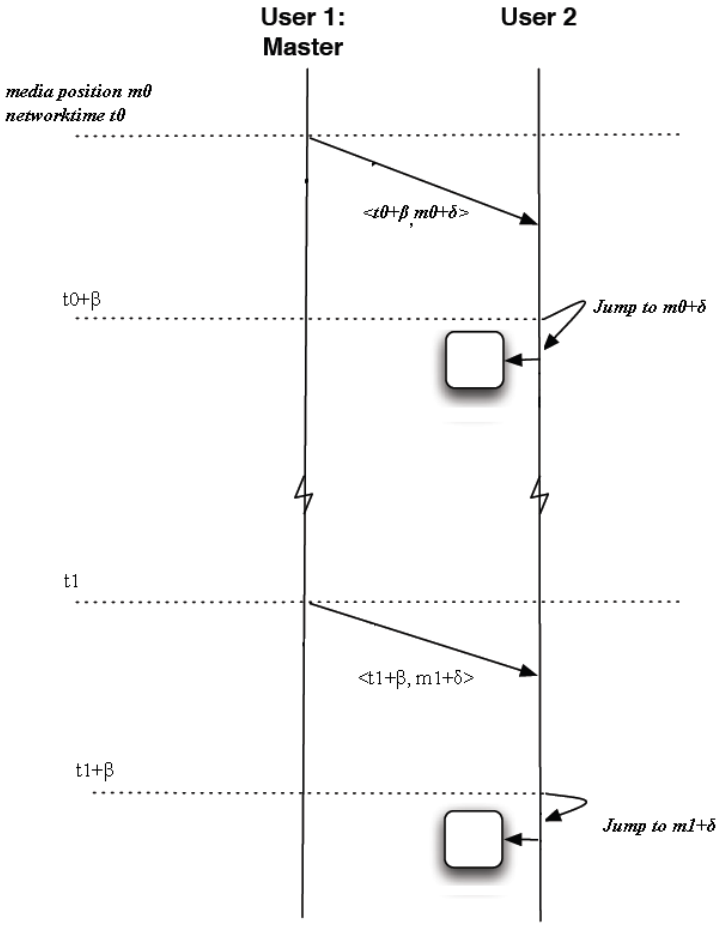


Figure 4.7 Illustration of the applied synchronization algorithm

The parameters δ and β were application/hardware specific and found using the validation setup from chapter 4 with a trial and error approach. Different parameters were tried for δ and β and by measuring with the setup from chapter 2 the right value was found for synchronization.

The system at the zero (synchronized) condition was recorded from the side for 3hrs. After segmentation and scene-change detection the cross-correlation function of the scenes described in 4.4 was computed. The result is shown in Figure 4.8 (similar to the 2.6). The peak was detected one frame later (k=1) compared to the fully synchronized case (k=0). This implies a difference of approximately 33 ms+- 16ms.

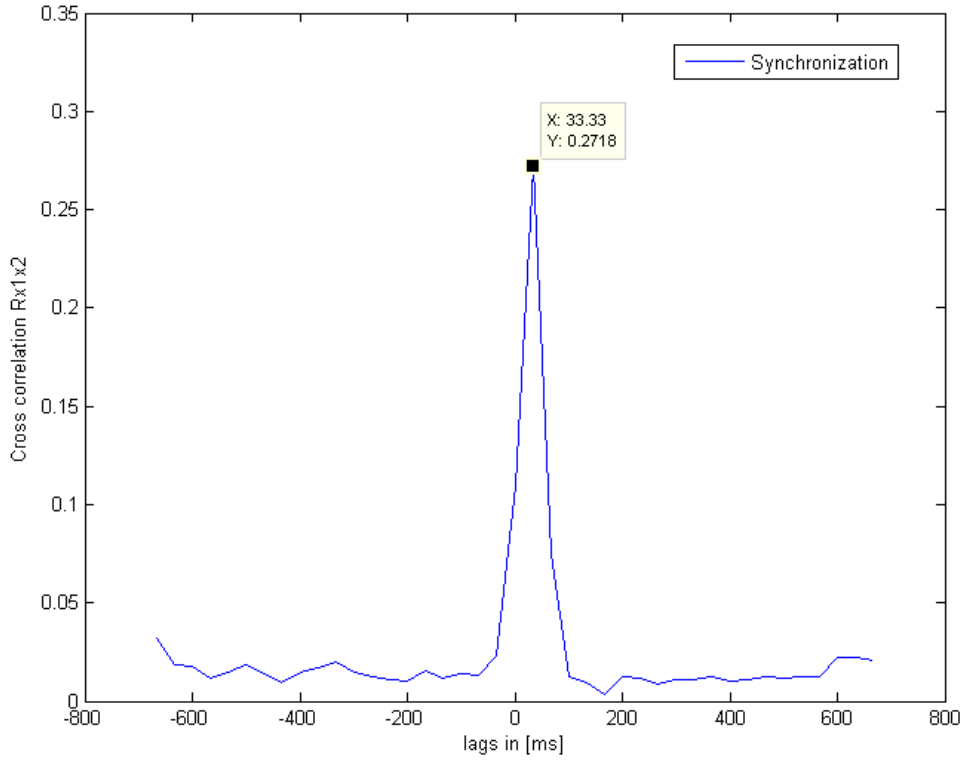


Figure 4.8 Result of validating the synchronized condition

System setting[ms]	0 ms
measurement [ms]	33.33

Table 4.7 Result of validating the synchronized condition

Also the correlation model was used on the measurements from the different synchronization conditions to validate the system at different synchronization conditions. Scene data from 9 samples of 15 minutes side-recorded were used and gave cross correlations from the measurement system that are shown in Figure 4.9. It can be seen from Figure 4.9 that the peaks correspond well to the applied system setting indicating correct operation of the synchronization algorithm. However the same small bias of the left (master) being ahead 1 frame is observed. The values of the system settings and the peak settings are compared in table 4.8. Table 4.8 shows that the left side (master) is always ahead by approximately one frame = 33 ms comparing to the system setting. This bias could not be eliminated from the system but is small enough for our setup to be considered valid for the user tests. Table 4.8 shows the measurement values in the different synchronization conditions.

System setting[ms]	**500**	**-500**	**1000**	**-1000**
System setting[frames]	**15**	**-15**	**30**	**-30**
Measurement[ms]	566.61	-466.62	1033.23	-966.6
System setting[ms]	**2000**	**-2000**	**4000**	**-4000**
System setting[frames]	**60**	**-60**	**120**	**-120**
Measurement[ms]	2033.13	-1966.47	4032.93	-3966

Table 4.8 Results of validating the out of sync conditions

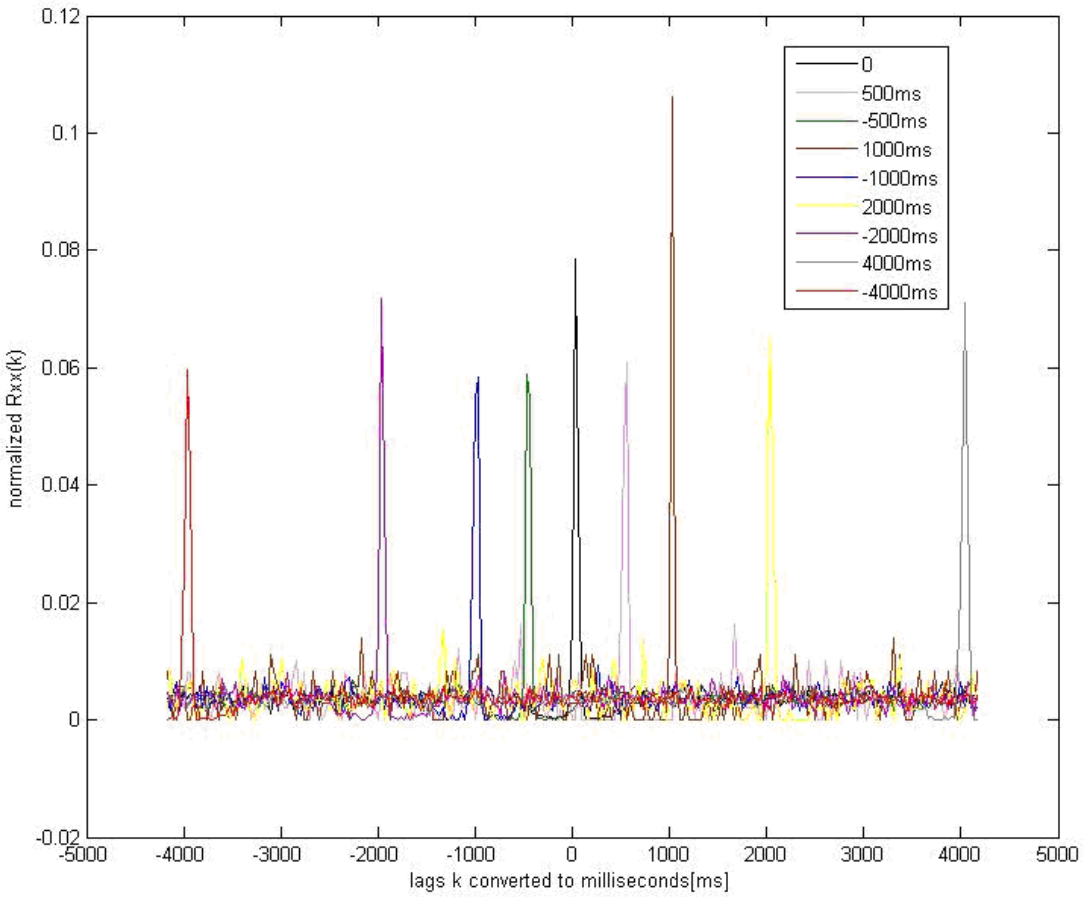


Figure 4.9 plot of the cross-correlation function between the measured scene-changes of device-1 and device-2 in experimental setup

4.6.2. Locations

The experiments were carried out at the KU Leuven CUO lab and spread over two locations. The main hall where participants were received and where the first participant watches his video is shown in Figure 4.10 while the second separated room is shown in Figure 4.11

Figure 4.10 The main room for user testing and receiving participants Figure 4.11 The second separated room

4.6.3 Equipment and software

Teamspeak software was used to facilitate voice and text chat over the Local Area Network. Teamspeak [44] software is an application supporting high quality VoIP and text chat and is often used by gamers. Also we used a sennheiser headset which clearly isolates the TV sounds coming from the headphone from the input microphone. A camera was used to record both participants for non-verbal behavior.

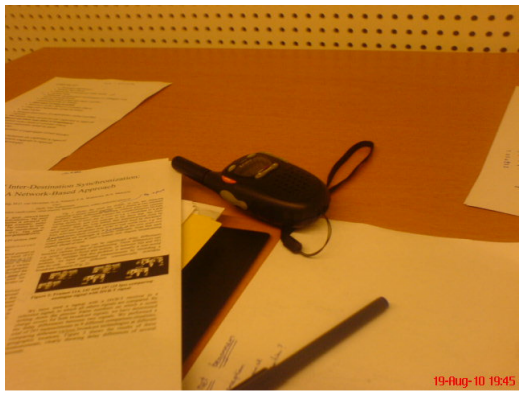


Figure 4.12 Sennheiser headset used for voice-chat and watching the video Figure 4.13 A handheld transceiver was used for communications between the administrators

Talk radios were used for the administrators to communicate with each other (mainly to announce when a questionnaire needed to be filled in, and when the video could be started). The control room contained infrastructure for recording and coding the events in our experiment. The microphone can be used to give instruction to the control room, the large PC contains NOLDUS software for coding and simultaneous video recording from different angles. A small laptop in the LAN contained the administrator software for the synchronization algorithm. With the equipment described in this chapter an environment was created to perform the test with actual users. The headsets and the PC's used correspond with ordinary pc and headsets people may use in a social TV situation.

4.7 The success of the test from preliminary test results

This section performs an evaluation of the designed and conducted experiment.
The actual tests were conducted from 23/8/2010 up to 29/8/2010. The first two days were used to setup the laptops, the camera's and the teamspeak software. It was important to control the volume of the headset and the media player software as to adjust them to each other. Also the pre-edited version of the quiz show to be used was converted to a suitable format (mpeg2) needed by the synchronization algorithm.
The local lag algorithm was calibrated and validated which was already discussed in section 4.6.
On 24/8 the pilot test was performed which resulted in 0.8 seconds being noticeable in the case of chat. This value led us to take the values 500ms, 1000ms, 2000ms and 4000ms for testing conditions. From 25/8/2010-29/8/2010 38 participants entered the experiment in couples. Only one couple was removed from the results due to lack of cooperation. Only one 7 minute session was disturbed by technical failure in the network. This session was also removed from the results.

As the test was designed to guarantee high likeability, togetherness/connectedness and high conversational activities this is validated. Observations and interviews after the experiment indicated that these achievements were met. After the collected data was preprocessed (to remove the randomized ordering) we performed some concrete analysis to show how successful the test was.

One of the appropriate criteria for choosing the genre and the specific show was that participants would like the show (section 4.6). Overall out of five possible levels ranging no negatives were chosen. This is quite exceptional that nobody from 36 test persons disliked the show. An exception was the couple which didn't participate properly and was therefore removed from the final results. This reduced the number of test persons from 38 to 36. The results for likeability are shown in Figure 4.14.

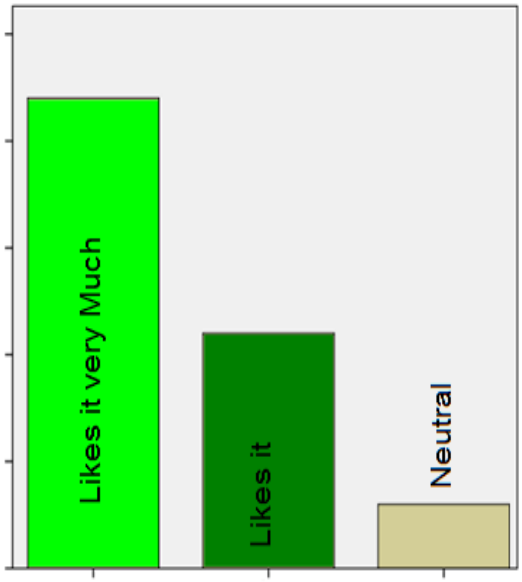


Figure 4.14 Liking of the show

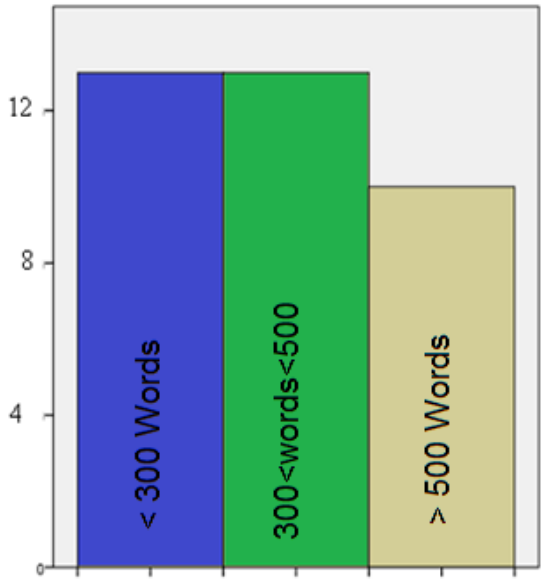


Figure 4.15 Chat activity in number of words per 35 minute session

Also data about the chat activity was analyzed. The chat logs of the sessions were stored by the teamspeak software used. We used a simple python script from pythoncsv to port the chatlog into a csv (comma separated value) file. On this csv file we used filters from Microsoft excel to separate the different sessions and different users. The results of the filtering were copied to Microsoft word version 2007 were the word and character count was obtained. The results were stored in the respective SPSS data files. The content of the chat was also inspected to involve around the show content. From some samples (100 lines of text) it was seen that 79 lines were directly or indirectly related to the show. The results of the chat activity are shown in Figure 4.16. The median number of words chatted was 372 corresponding to approximately 10 words per minute. An analyzed sample of the chatlog showed that 80% of chat was centered around the content of the show.

To measure if users felt together/connected six questions were posed after each condition. The questions are shown below. The answers are from 1 (strongly agree) to 7 strongly disagree.

1. The contact with my conversation partner was superficial (1-7)
2. I felt together with my conversation partner (1-7)
3. I felt that my conversation partner did not understand me well (1-7)
4. I felt connected with my conversation partner (1-7)
5. I got little satisfaction out of the conversation (1-7)
6. I felt that my conversation partner and I could talk well to each other (1-7)

As can be seen questions 2, 4 and 6 are positively formed while 1,3 and 5 are negatively formed. Therefore we rescaled questions 2,4 and 6 making all questions negatively formed with the goal to calculate the consistency of the answers. In the case 7 (strongly disagree) is than the highest value for togetherness. The 6 questions are taken together and tested for reliability. A usual method from social science to test if responses on questions are reliable (consistent) is done with calculating a correlation coefficient. The results gave Cronbach's alpha=0.852 and a Gutmann's split half=0.807. In social science values above 0.7

are usually seen as reliable/correlating (i.e. measuring the same thing). After this we took the (rescaled) questions together and determined the average. This average is from now on used as a measure for the togetherness/connectedness. The 95 percentage confidence regions of the average togetherness of voice and chat are shown in Figure 4.16. The results confirm the previously published hypothesis [36]that text chatters feel less together than voice chatters. For our user test however, the most important results is that high levels of togetherness are achieved for both cases. Voice chatters feel from somewhat together up to together on average while text chatters feel from neutral to somewhat together.

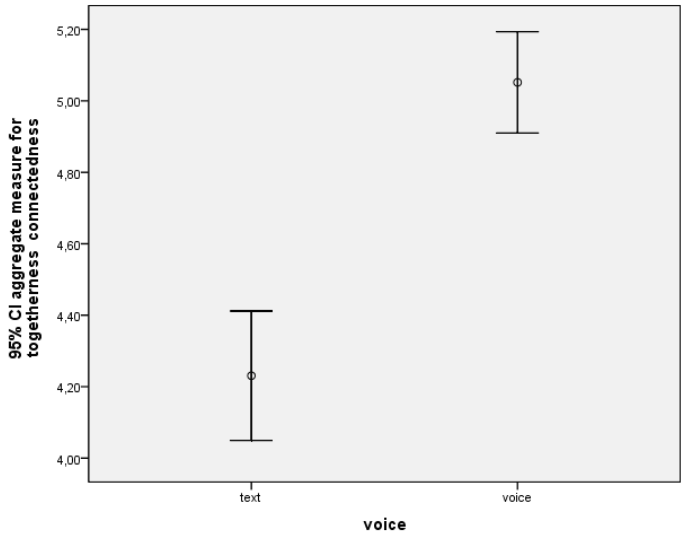


Figure 4.16 Togetherness and synchronization level, especially the voice chatters experience a high level of togetherness connectedness

From this assessment we can draw the following conclusions regarding the designed and performed user test.

1. *The user test was technically successful despite one minor technical network outage in one 7 min session. Especially as a previous test preformed by KU Leuven and CWI completely failed technically*
2. *Togetherness/Connectedness was consistently measured*
3. *High likeability, the participant knowing the show, togetherness and conversational activity were all achieved, making the test successful from a research perspective.*
4. *Conversation was structured mostly around the content of the show, as desired*

The user experience is considered successful as both technical failures have been largely avoided. And the user experience requirements from this design have been largely met.

4.8 Conclusions/Future work

This chapter presents an experiment design that was never run before. From Social TV research and telecommunications standards the two most relevant constructs were used to develop research questions, questionnaires and hypotheses. A proper use case was analyzed. The use case chosen was confirmed by the first experimental results to be likeable and enhance social interaction. Validity threats such as recruitment, habituation effects were systematically assessed and eliminated. Correct test conditions were chosen based on pilots. By testing each condition on each participant, test results can be analyzed taking differences between participants into account. A technical setup for the experiment was presented and validated for its performance with the system from chapter 2. The first results show consistent high togetherness and likeability of the show experienced by the participants.
The experiment developed here can be re-used by other social TV developers and HCI researchers to do a similar experiment. For example the randomized scheme, use case, technical setup can be reused in other tests.

5. Statistical Analysis results of user test: The effect of IDMS on Perception and Annoyance

This chapter uses the results from the user test to answer research question 1, how much play-out difference do users notice and how much play-out difference makes them feel annoyed.

5.1 Perception and Annoyance instead of MOS

One of the main aims of the experiment was to find the relation between quality of experience (MOS) and IDMS. As can be seen in Figure 5.1 MOS 5 was very dominant in the results. The average (IDMS) MOS for text chat was 4.1061 and for voice it was 4.428. For investigating the effect of IDMS on MOS score the range of selected values for the synchronization difference is simply not well enough spread. MOS score 1(blue) is hardly selected, and approximately an equal number of times in each sync condition for voice chatters. We looked at when the synchronization difference becomes perceptible and annoying instead and aimed to find play-out differences that can be considered as threshold.

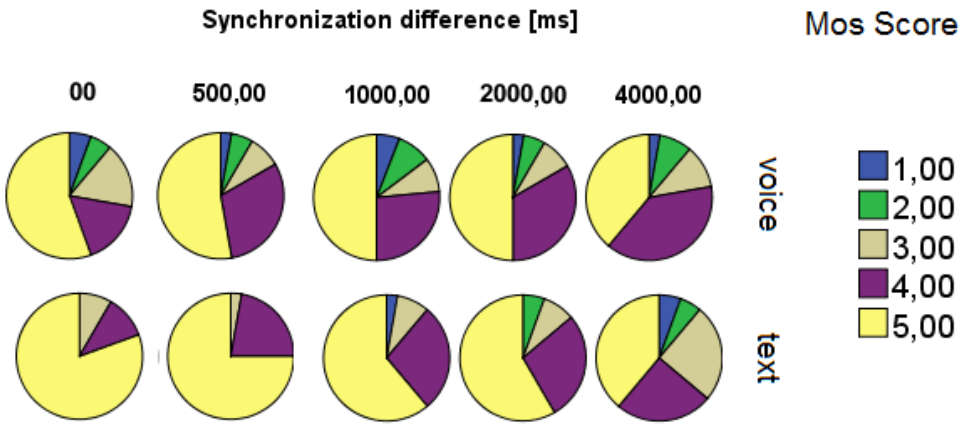


Figure 5.1 The spread of the MOS values resulting from the user experiment

Looking at Figure 5.2 we see that the dataset looks random for text chatters. Text chatters claim to notice a synchronization difference 50% of the time while being fully synchronized. This can be merely explained due to perception of delay in typing as no perceptible delay in the chat software was perceived when comparing both screens. Also about 40% of voice and text chatters being out of sync with 4s do not notice synchronization differences. In the voice chat case results show more of the expected trend as synchronized voice chatters have a low percentage of "incorrectly" noticing.
The graphs show that for voice chat a clearly increasing trend for annoyance and perception is observed while for text chat this was not the case. This leads us to form two new hypotheses:

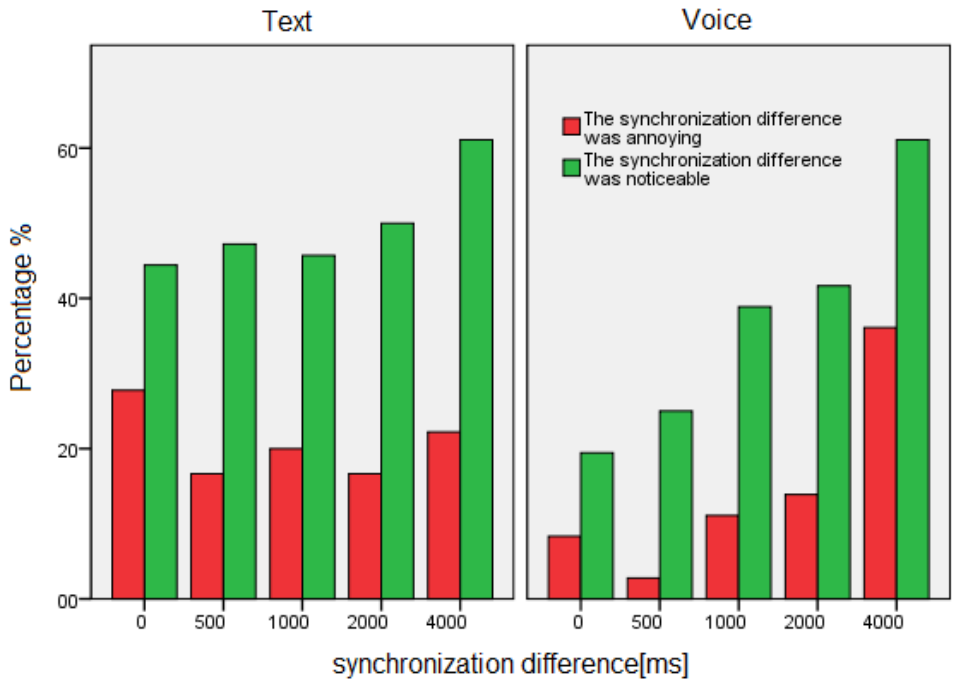


Figure 5.2 Bar chart representing percentage of participants annoyed/noticed in each condition

1. Voice chatters notice differences in the 0-4s range and may get annoyed by it

2. Text chatters did not notice synchronization differences or get annoyed by it comparing to the synchronized condition

To accept or reject these hypotheses we aim to use appropriate statistical methods. In case of hypothesis 1 we would like to also find the "thresholds" of annoyance and perception.

5.2 Statistical methods for analysis

Two statistical methods for testing thresholds/differences from literature are presented. In section 5.1. we present the psychometric methods which are often employed in perception assessment of audio/video applications. Second a non-parametric test for finding difference between related samples is presented in section 5.2. Based on the dataset and specific characteristics of the methods, the best one is selected for analysis.

5.2.1 Psychometric methods for finding perception thresholds

Psychometric methods were introduced by Fechner in (1860/ 1966) in his book "Elementen der psychophysik" [45]. The main topic of psychophysics defined by Fechner was to study the relation between the physical world and the phenomenal (perceived/experienced) world (perception by people). The concept is illustrated graphically in Figure 5.3. The cyan box refers to the physical world in which a stimulus (play-out difference) occurs. The green box refer to the sensory organs/neurons that perceive the physical signal. The red box is the perceptual/conscious perception perceived by the person. The relations between these "boxes" are studied by psychophysicists.

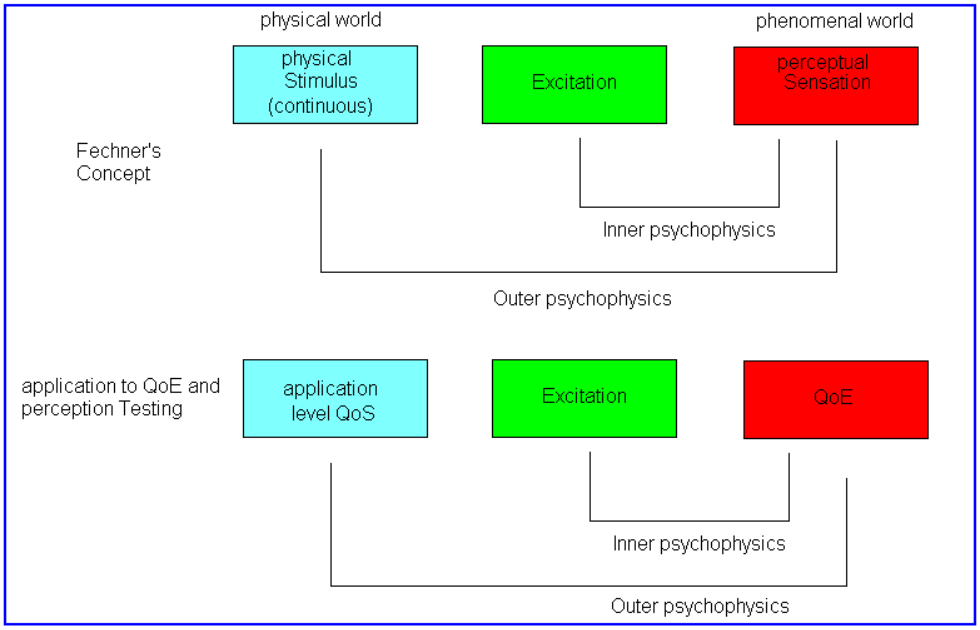


Figure 5.3 Fechner's concept of psychophysics which studies relation between physical and experienced world

Outer psychophysics studies the relation between (physical) real world stimuli and the (objective) perception. Inner psychophysics study the relation between neural stimuli as for example present in the brain and the perception/sensation.

Another good and relevant question to our research objective is how do psychophysicists find this relation between physical stimuli and "objective" experience? Objective experience seems strange as experience can be different amongst people and are therefore subjective.
The answer to this question is given in the specific sub-area of threshold psychophysics which is used to find thresholds. Three methods in classical psychophysics are often employed: constant stimuli, method of limits and the method of adjustments. All three methods will be explained briefly. For more rigorous explanations the reader is referred to [45].

All methods are based on exposing a stimulus to a test participant who communicates his/her sensation followed by a change in the stimulus intensity This procedure is usually repeated until a difference in sensation is communicated by the participant.

The method of adjustment allows participants themselves to adjust the stimulus until the stimulus becomes just noticeable/ just not noticeable. In general both ascending and descending series are tried as doing only one direction is known as a possible bias of the result. Another method is the method of limits in which the experimenter controls the intensity of the stimulus increasing/decreasing towards the threshold, the test participant can at each step indicate if a difference or stimulus is observed. Often a staircase pattern is often employed. In the staircase pattern, after a difference in sensation is detected in the ascending direction the test is repeated in the descending direction until the difference in sensation is detected. This is repeated until for both the descending and ascending direction the same stimulus value is found. This value is the threshold found using the method of limits. Another method is the method of constant stimuli were the participant is exposed to many stimuli in randomized order. The responses are then plotted and fitted (often to a normal curve). The 50% level us usually taken as threshold in this method and the region between 25% and 75% as the region of uncertainty.

Threshold finding using psychometric methods is accurate and well founded. Unfortunately all methods require many measurements on the test participants. This is not a problem when measurements can be performed quickly as in hearing or vision tests. In our social TV test one sample takes at least 7 minutes. Due to this lack of data it is difficult to apply psychometric methods.

5.2.2 Non-Parametric methods for finding effects in perception test

The non parametric Cochrane-Q tests have a lot of similarity with repeated measures ANOVA as they take within effects into account. Cochrane's Q tests are used for binary and discrete variables. The methods are non-parametric, i.e. they don't assume any prior distribution. The test can detect differences between k related samples. The details of the calculation of the test statistic can be found in [46] and the statistical supplement on the DVD. The conceptual model of the Cochrane Q-tests is given below in Figure 5.19. By comparing the synchronized condition with each other condition we can find which synchronization differences are perceived significantly and which ones are not.

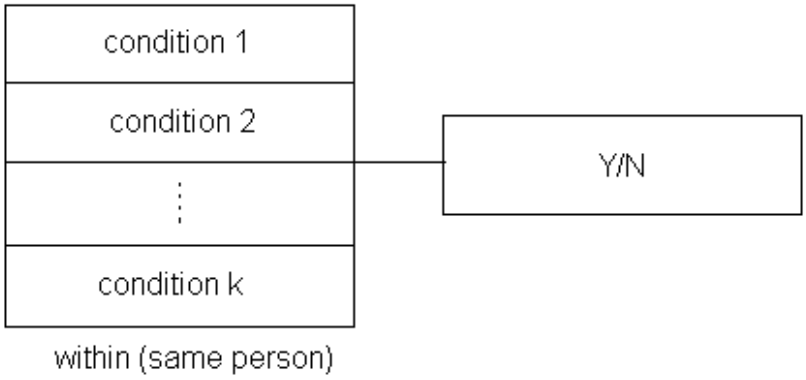


Figure 5.4 Cochrane's Q test

5.2 Applying Cochran's Q to test perception and annoyance

In this section we apply Cochrane's Q test to compare differences in annoyance/perception between the synchronized condition and the condition with play-out difference. The test is repeated for different groups: voice and text, seen before and not seen before and chat activity/experience. These factors were found in the exploratory phase to have an effect on annoyance/perception of play-out difference. By applying Cochrane's Q test we take variance between different participants better into account than when using psychometric methods. With Cochrane's Q test we also take the binary nature of the response into account, which is not the case for t-test's or ANOVA's. Also by using the synchronized condition as a base for each comparison, we carefully find the effect of play-out difference for each condition.

5.3.1 Text vs Voice chat

For voice-chat both significant differences between synchronization conditions and perception as between synchronization difference and annoyance are found. The results of comparing synchronized with non synchronized condition with Cochrane's test are in Table 5.1 . The conditions V500, V1000, V2000 and V4000 correspond to the amount of difference in ms compared to being synchronized with voice. The T500 to T4000 conditions are comparisons to the synchronized condition for text-chat.
The differences detected as significant are marked green while the non-significant differences are not marked. The results show that text-chatters do not notice or get annoyed significantly. Voice chatter notice 1s and more and get annoyed with 4s of difference.

Cochrane's test comparing perception and annoyance to the synchronized condition									
		V500	V1000	V2000	V4000	T500	T1000	T2000	T4000
Perception	Q	0,667	7,00	5,33	11,845	0,333	0,692	1,6	2,250
	p	0,414	0,008	0,021	0,001	0,564	0,405	0,206	0,134
Annoyance	Q	2,00	0,200	1,00	7,143	2,00	0,00	2,00	0,33
	p	0,157	0,655	0,317	0,008	0,157	1,000	0,157	0,564

Table 5.1 Results of voice and text chat, green boxes indicate significant results

Conclusion 1a: *The results of the experiment show that text chatters do not notice synchronization or get annoyed by it compared to being synchronized. Voice chatters on the other hand perceive synchronization difference significantly at 1 second or more and get annoyed at 4 seconds or more.*

5.3.2 Having Seen the episode Before

From Figure 5.5 the fact that the show was seen before or not seems to make a large difference in perceiving synchronization difference or getting annoyed by it. N=16 people saw the episode before while the other N=18 people did not see the episode before. A Cochrane test on both subsets of the data was performed to see if significant differences existed. Again for chatters no significant differences were found. For voice chatters the tests concluded that people who saw the episode before notice differences significantly, while first time viewers did not.

Cochrane's test comparing seen before and not seen before for Voice condition									
		Seen Before				Not Seen Before			
		V500	V1000	V2000	V4000	V500	V1000	V2000	V4000
Perception	Q	0	5	6	13	0,66	2	0,66	0,66
	p	1	0,025	0,014	0,0001	0,414	0,157	0,414	0,414
Annoyance	Q	X	3	3	10	X	2	1	0
	p	X	0,655	0,317	0,002	X	0,157	0,317	1

Table 5.2 Results of Q test based on seen before/not seen before

Conclusion 1b: *Having seen the episode before makes detecting IDMS much easier for the voice case. Voice chatters that have seen the episode before notice synchronization difference, participants that did not see the episode before do not.*

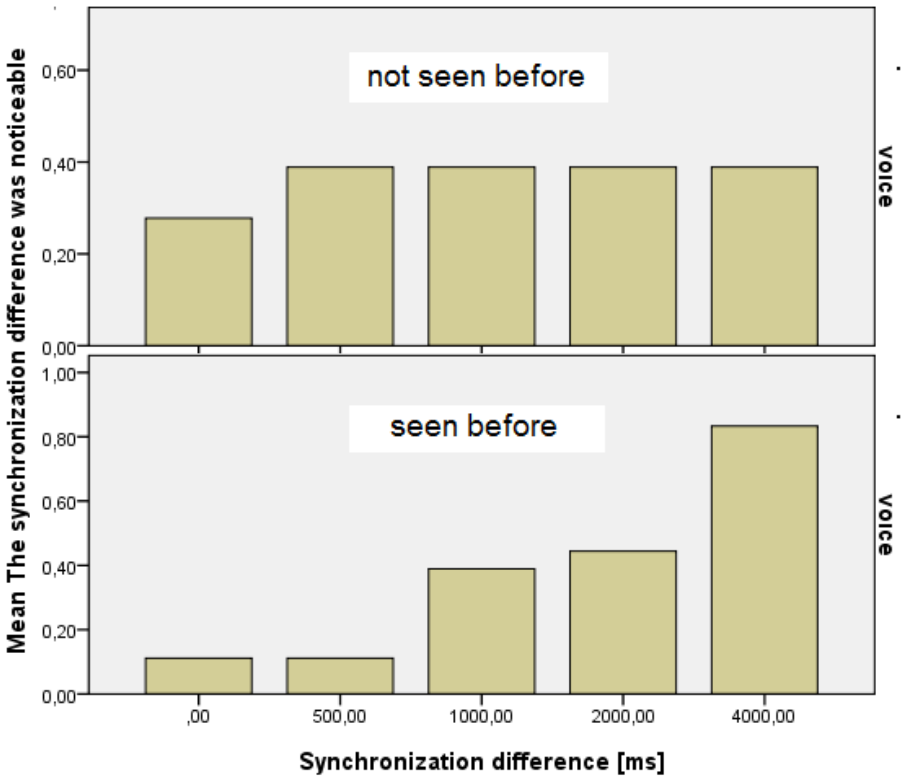


Figure 5.5 Data based on having seen the episode before

5.3.4 Chat activity/experience

Chat experience of the participants was obtained in a questionnaire after the test. It ranges from never to everyday. The chat experience was found to have no significant effect on the noticeability/annoyance of synchronization.

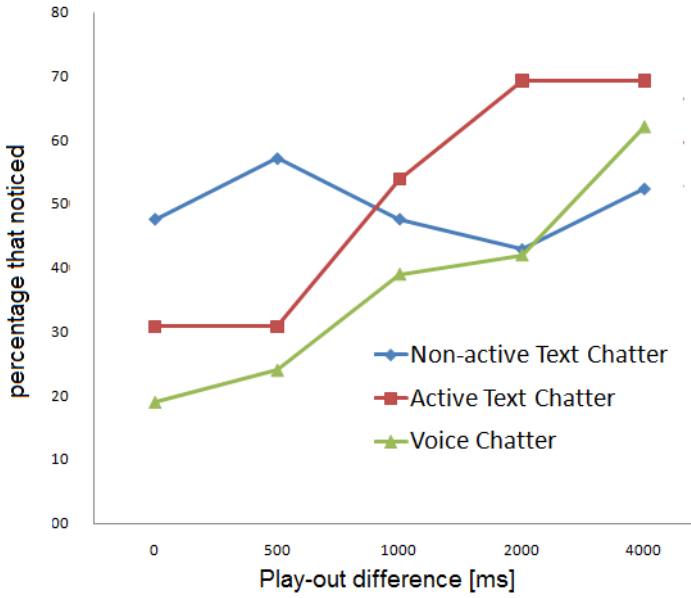


Figure 5.6 Percentage noticed and chat activity

On the chat log data we performed a word count and a character count to quantify the chat activity of participants. From the data the group of participants was split in an active group (N=15) that typed more **Conclusion 1c:** Non active text chatters are in general not able to detect synchronization differences. Active

chatters however notice synchronization differences but not as strongly as voice chatters (they notice only 2s or more). Cochrane's Q-test showed that active text chatters do notice synchronization differences from 2 or 4s compared to being synchronized based on a 5% significance level.
than 400 words and a non active group (N=21) that typed less than 400 words. Differences in notice ability between the two groups and voice chatters are shown in Figure 5.6. The behavior might look strange as it based only on 5 synchronization conditions. In general the data contained a lot of randomness with negligible effect of conditions and external factors on IDMS. However the effect of chat activity is clearly shown and also tested to be significant by using Cochrane's Q test.

To test the if active chatters indeed notice play-out difference significantly a Cochrane's test was performed on the set of the active chatters and the set of the non-active chatters. Table 5.3 shows the results of comparing each condition to the synchronized condition using Cochrane's Q test. The results show that active text chatters that use more than 400 words notice play-out difference significantly.

Cochrane's test comparing Active and Non active chat conditions									
		Active Chat				Non active chat			
		T500	T1000	T2000	T4000	T500	T1000	T2000	T4000
Perception	Q	0,400	3	5	5	0,66	2	0,66	0,66
	p	1	0,083	0,025	0,025	0,414	0,157	0,414	0,414
Annoyance	Q	0	3	1	3	X	2	1	0
	p	1	0,083	0,317	0,083	X	0,157	0,317	1

Table 5.3 Cochrane´s Q test for active and non active chatters

Conclusion 1c: *Non active text chatters are in general not able to detect synchronization differences. Active chatters however notice synchronization differences but not as strongly as voice chatters (they notice only 2s or more). Cochrane's Q-test showed that active text chatters do notice synchronization differences from 2 or 4s compared to being synchronized based on a 5% significance level.*

5.3.5 Lagging ahead/Being behind

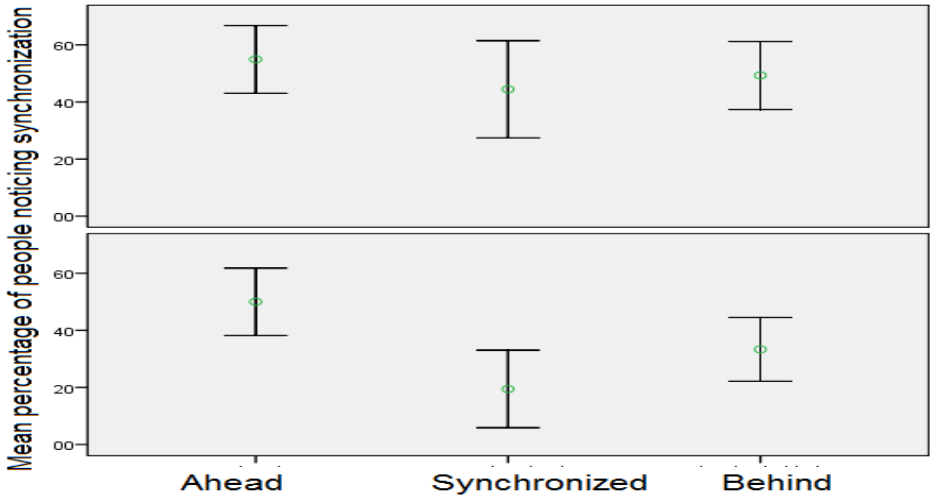


Figure 5.7 The effect of being ahead-behind

In the experiment, on average some small difference in notice ability and annoyance was found between lagging participants and leading participants. This small effect was seen in both the voice and text condition. We did not test each lag/leading condition on each participant; therefore we cannot use our analysis method to show significance of these results. The effect needs further investigation. Figure 5.7 shows error bars around annoyance and perception. It shows that on average people ahead notice the difference more. This is both the case for voice chat in the upper column as for text chat (lower column).

5.4 Conclusions

The results of the experiment show that voice chatters notice play-out difference and can get annoyed by it. However if groups of chatters are taken into account it turns out that active chatters with more than 400 words per session also notice play-out differences of 2 and 4 seconds. Taking both active text chatters and voice chatters into account a play-out difference bound of less than 1s seems a good recommendation for social TV. Voice chatters that had seen the episode before may notice it, but in general play-out differences below this bound will go unnoticed.

6. Statistical Analysis results of user test: The effect of IDMS on the Social Experience

The social experience was derived from 6 questions that measured the feeling of togetherness experienced by participants. It was shown that these questions were consistently answered (Cronbach's alpha=0.852 and a Gutmann's split half=0.807). The effect of play-out difference on the feeling of togetherness is analyzed statistically in this section. First we propose the appropriate statistical methods, followed by a section 6.2 that presents the results of running the test on the experiment data.

6.1 Statistical analysis methods

6.1.1 Repeated Measures ANOVA

The difference in togetherness between synchronization conditions will be tested using repeated measures analysis of variance. This is a technique similar to one-way ANOVA that also takes dependencies of responses by the same participant into the account. The mathematical details of the method are explained in [46] and the statistical supplement on the DVD. Our experimental design shown in Figure 6.1 is a typical repeated measures design as conditions are measured on each of the participants. The repeated measures test requires normality of the underlying variables and sphericity. With the statistical software package SPPS it is possible to test if both of these conditions are satisfied.

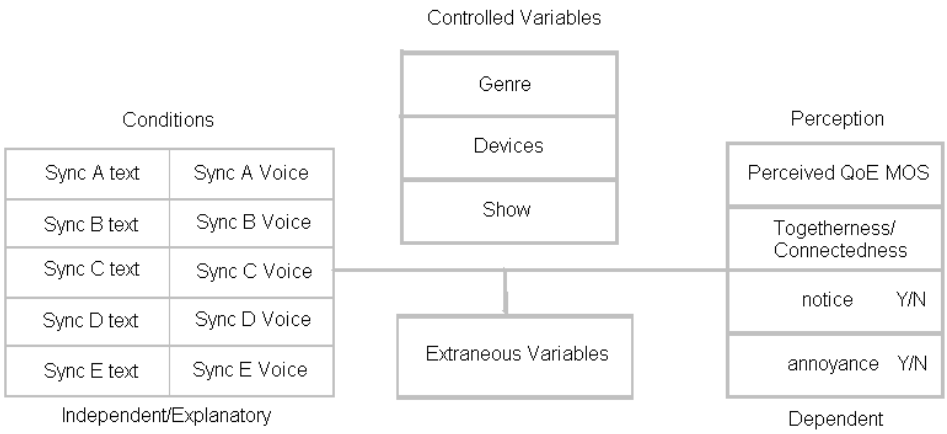


Figure 6.1 The experimental design, a typical case of repeated measures

6.1.2 Paired samples t-test

An appropriate statistical method to prove the difference between the voice and chat condition is the paired samples t-test. This test compares two related samples and checks whether the mean difference is significantly bigger than zero. As we are explicitly interested in the difference between chat and voice we compare each paired (voice, text) sample *in the same synchronization condition* and *for the same participant*. We had to alter our dataset to perform the analysis in the SPSS software that was used. We did this by changing the rows based on participant and synchronization condition. The paired t-test does not need equal variance like one-way ANOVA or independent samples t-test. This is useful as the variance amongst text chatters was bigger than amongst voice chatters.
Also the pair's relation (same synchronization condition and same participants) is also taken into account making the paired samples t-test ideally suited to compare text-chat with voice chat. The details of the method are given in [46] and the statistical supplement on the DVD.

6.2 Results

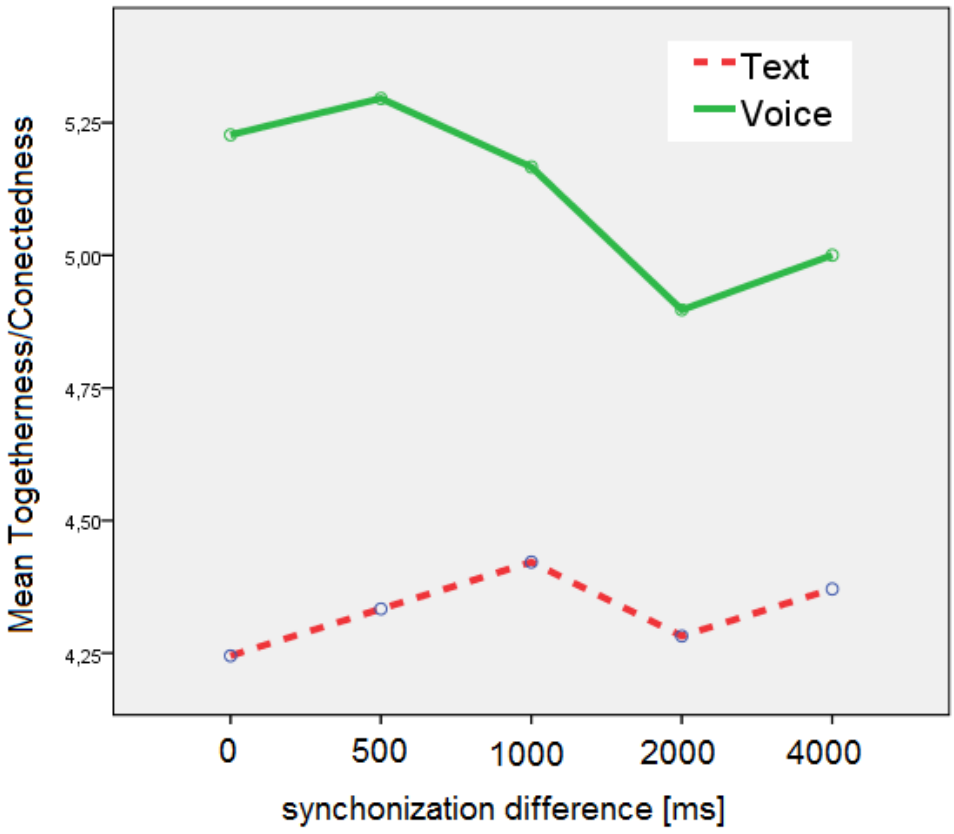


Figure 6.2 Togetherness levels in TV watching experiment

Figure 6.2 clearly shows the differences in togetherness levels between voice and text conditions. To test this difference statistically we used the paired sampled t-test. Samples were paired together for the same participant and the same synchronization condition.

Participant 4	Sync 500	text	Participant 4	Sync 500	voice
Togetherness level 4.3			Togetherness Level 5.1		

Table 6.1 A paired togetherness example

After normality assumptions were verified, the result of the paired samples t-test gave an average difference in the togetherness of 0.78 points with a standard deviation of 0.11. The t statistic is was $t(179)=-7{,}143$ $p<0.001$. This indicates that if togetherness/connectedness levels were equal for voice and text chat the probability of an outcome with differences as big as observed in our experiment was less than 0.1%, therefore we conclude a significant difference between the voice and chat conditions.

To test the effect of synchronization condition combined on togetherness repeated measures ANOVA was performed on the data set.
The resulting test was successful in that both necessary assumptions for repeated measures ANOVA normality and sphericity were met for both text and voice condition which were tested separately. For the togetherness in the text chat condition no significant difference was found for synchronization conditions $F(4,140)=0.33$ $p>>0.05$. For voice chat the test gave significant difference $F(4,140)=2.989$ $p<0.05$. The effect size of approximately 8% and the fact that only quadratic and fourth order fits model the contrasts while a linear model was expected, leads us to reject the hypothesis that play-out difference has an effect on the overall social experience/togetherness.

Conclusion 2a: *For both voice and text chat, the results show that play-out difference does not have a significant effect on the togetherness/connectedness experienced by the participants. Between the voice and chat conditions a difference in togetherness and connectedness exists*

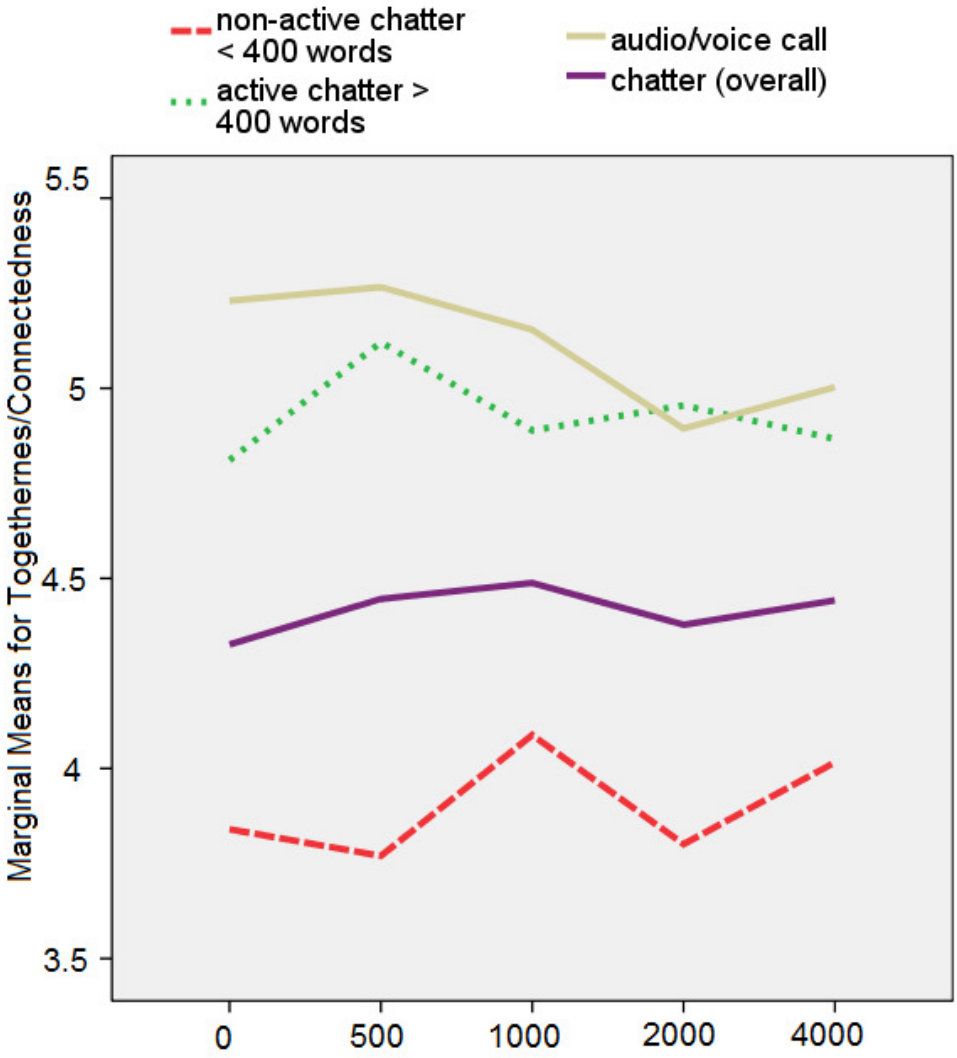


Figure 6.3 Togetherness/Connectedness based in experiment based on chat activity

Chat experience, having seen the episode before or likeability of the show were found not to have an effect on the togetherness level. In Figure 6.3 it is shown that the more active chat group (>400 words) experiences higher togetherness levels than the non active group.

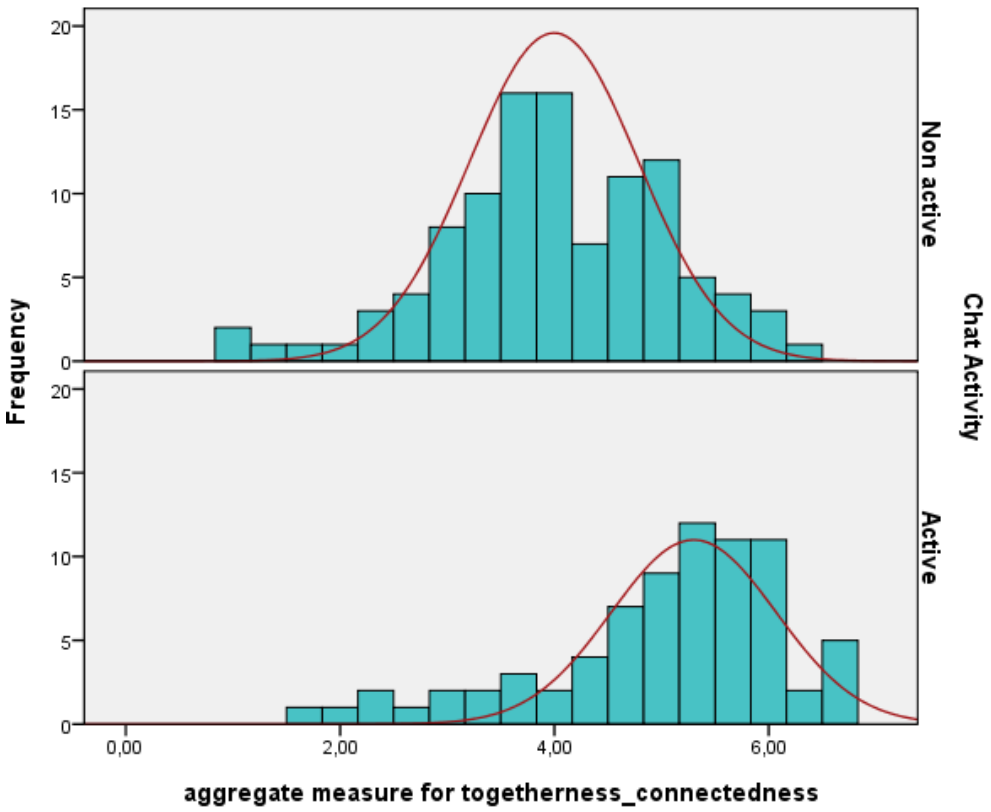


Figure 6.4 Histograms for togetherness, showing approximate normality

As both groups (active and non active chatters) showed to approximately normally distributed togetherness with approximately equal variance as shown in the histogram in Figure 6.4 we can use independent samples t-testing. This assumption was also tested to be valid in SPSS with a test for normality.
Testing for difference with independent samples t test in means between active and non-active chatters gave $t(178)=-6,2$ $p<0.001$ and a test for equal variance (Levene's) showed that variance was indeed approximately equal.

As significant differences between active and non-active text chatters were found, the next step was to compare these two groups to voice chatters. All three groups were already shown to be approximately normally distributed so we performed one-way ANOVA to test the differences between the groups. The test gave significant differences between the groups $F(2,357)=50.88$ $p<0.05$ and homogeneity of variance(approximately equal variance) between the groups. A consequently performed post- hoc test (a test comparing differences between the specific conditions) that checks the differences between each corresponding pair of groups showed that the difference between voice chat and active text chat was no longer significant. Therefore we can conclude that the overall difference in togetherness between voice and text chatters is mainly attributed to the less active text chatters. The mean togetherness was 3.9 for non-active chatters 4.9 for active chatters and 5.1 for voice chatters.

Conclusion 2b: *Difference in togetherness between voice and text chat is mainly attributed to non-active text chatters typing less than 400 words per second. Active chatters and voice chatters feel approximately equally more together /connected compared to non-active text chatters*

6.3 Conclusions

In this section the results of the user experiment performed in Leuven were analyzed using appropriate statistical methods. The results show that for the test case of couples play-out difference does not have an

effect on the togetherness/connectedness experienced. This result contradicts assumptions from earlier research such as [12] who suggested that inter-destination media synchronization for shared TV watching enhances the social experience. For social TV design the group of active chatters/ voice chatters can be an important target group as they experience high togetherness. This group notices play-out difference significantly as shown in chapter 5, making inter-destination synchronization difference desirable to guarantee the best perceived Quality of Experience.

7 The Soccer Watching experience

This chapter investigates the user experience of play-out differences when watching a soccer match. It studies two participants that watch football together connected through an audio link. This can represent a social TV application or neighbors in vicinity of each other hearing each other cheering. A study of this type of study wasn't run before. The results are compared for differences to the results from the quiz show in chapter 5. Compared experiment from chapter 3 we will only look at perception and annoyance aspects and an audio link. The results of this study are mainly useful to advice service providers of TV channels on the effect of IDMS.

7.1 Problem description

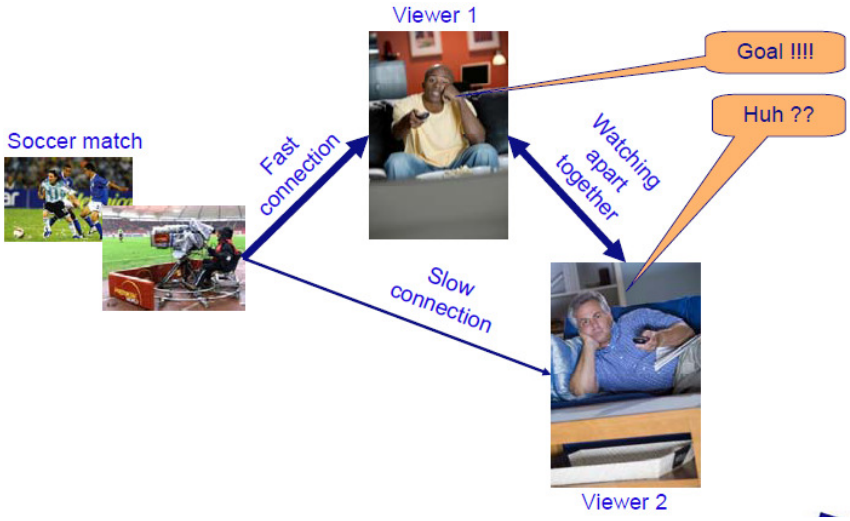


Figure 7.1 The effect of the inter-destination synchronization problem in a football match [47]

Figure 7.1 shows the typical use case, viewer 1 sees the goal while viewer 2 sees nothing. The cheering by viewer 1 can spoil the experience of viewer 2. In chapter 6 we showed that in the case of a sociable quiz genre difference of 1 second and higher become noticeable and 4s becomes annoying. The main question answered in this chapter is: "how much difference becomes annoying or noticeable?". We are also interested to see if participants are willing to change their provider for watching football. As the 10 recruited soccer fans reported that they liked football and watched regularly we assume to have a representative sample of soccer fans.

7.2 Technical Setup

In this experiment available resources at the TNO work environment were utilized to obtain an experimental setup. The setup features two office Cisco phone's with speaker option, two laptops for watching video and a control PC. Figure 7.2 shows the experimental setup. The connected laptop PC's and phone's are located in distant rooms. The administrator is located in a third room (or one of the other two rooms) and has the ability to control the video on both laptops located near each participant. This control functionality was implemented by using the VLC (videolan.org) http interface. A small computer program developed in java simultaneously calls the two video players to obtain synchronization, or with a fixed difference to obtain fixed offsets. The principle of this synchronization was discovered in earlier laboratory experiments from chapter 3. The stability of the VLC media player causes two similar video files played on two different PC's to remain synchronized if started simultaneously. To verify synchronization levels the experimental setup was tested using the same video files and the measurement tool from chapter 2.

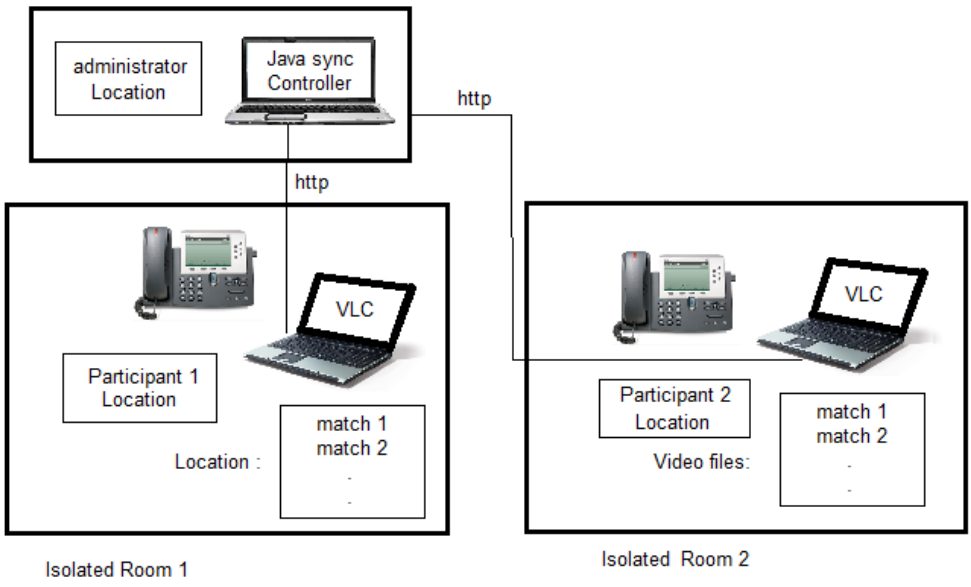


Figure 7.2 The experimental setup: a Cisco phone and a laptop running VLC

The resulting cross correlation functions of the validation are shown in Figure 7.3. An average absolute difference to the system time of approximately 60ms was found. We assume for the experiment that this synchronization level is sufficient. The response is not symmetrical and not always completely similar, The java controller calls the PC in room 1 using the VLC http interface first, and immediately continues with calling the PC in the second room to start the video. In practice this results in somewhat diminished play-out difference setup accuracy. However as Figure 7.3 shows the inaccuracy is kept below 100 ms which we considered appropriate for this specific experiment.

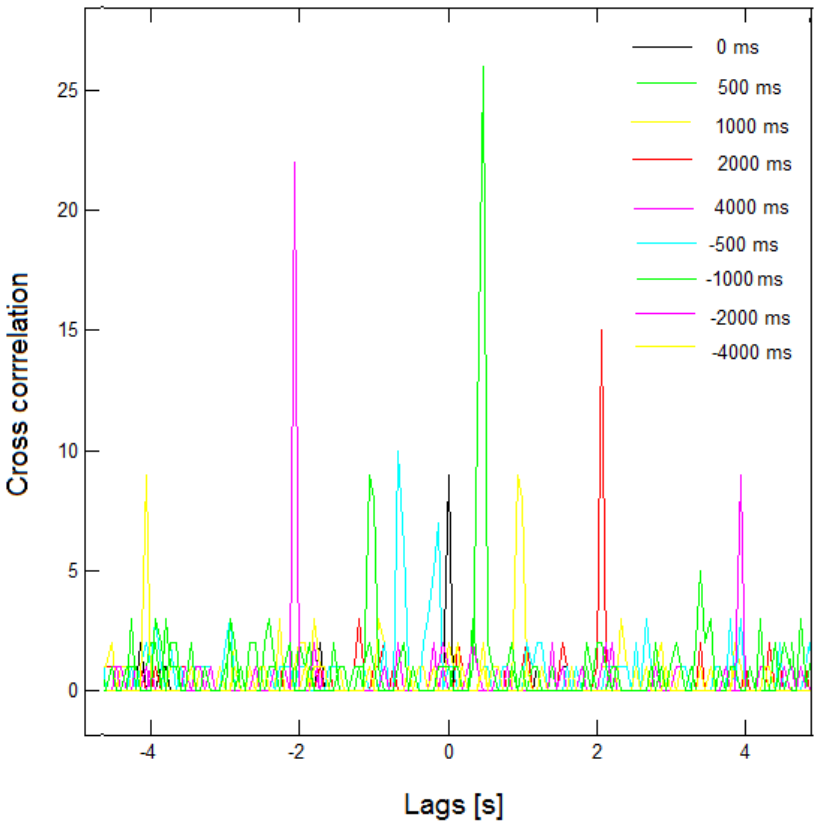


Figure 7.3 The cross correlation synchronization plot from the experimental setup using the measurement system developed in chapter 2

7.3 Use case / Questionnaire

We present the user with summaries from Champions league matches from 1995. For each session we present either to summaries of two matches of approximately 4 minutes, or one longer segment of 9 to 11 minutes. The first three sessions consisted of two short summaries while the last two sections consisted of the longer sessions. In each session a different synchronization level is applied in a similar random order as in table 4.5. to eliminate habituation effects. The same question as in 4.6 is asked. After this question the user is asked multiple choice if he would like to change the provider (Yes, No, maybe).

7.4 Results

Similar as in the social TV experiment from chapter 4 the results show that the high QoE scores are dominant. However it is less the case as in chapter 6 as in the case of 4 seconds of difference more than 50% of the participants gets annoyed.

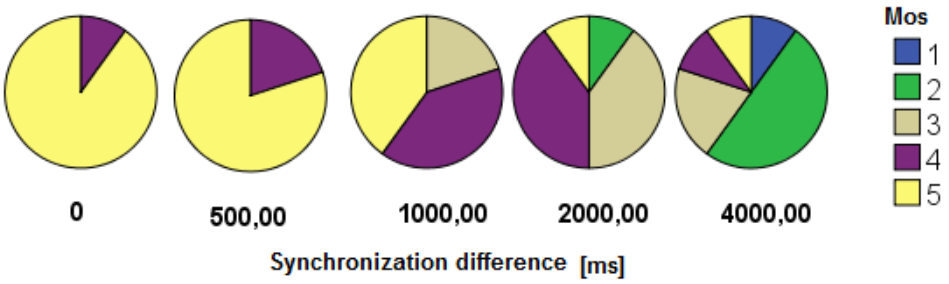


Figure 7.4 MOS values obtained in soccer watching experiment

Again we will look at the annoyance and notice ability to compare with the social TV use case. The results for annoyance and notice ability are shown in Figure 7.5 and show a similar trend as the voice chatters in the social TV experiment. 1s difference seems critical for noticing while 4 seconds is clearly annoying. Due to the small sample size of 10 persons we refrain from further statistical analysis. While this experiment with 10 participants is to small on its own to draw general conclusions, together with the experiment from chapter 3 it gives insights on the perception of IDMS.

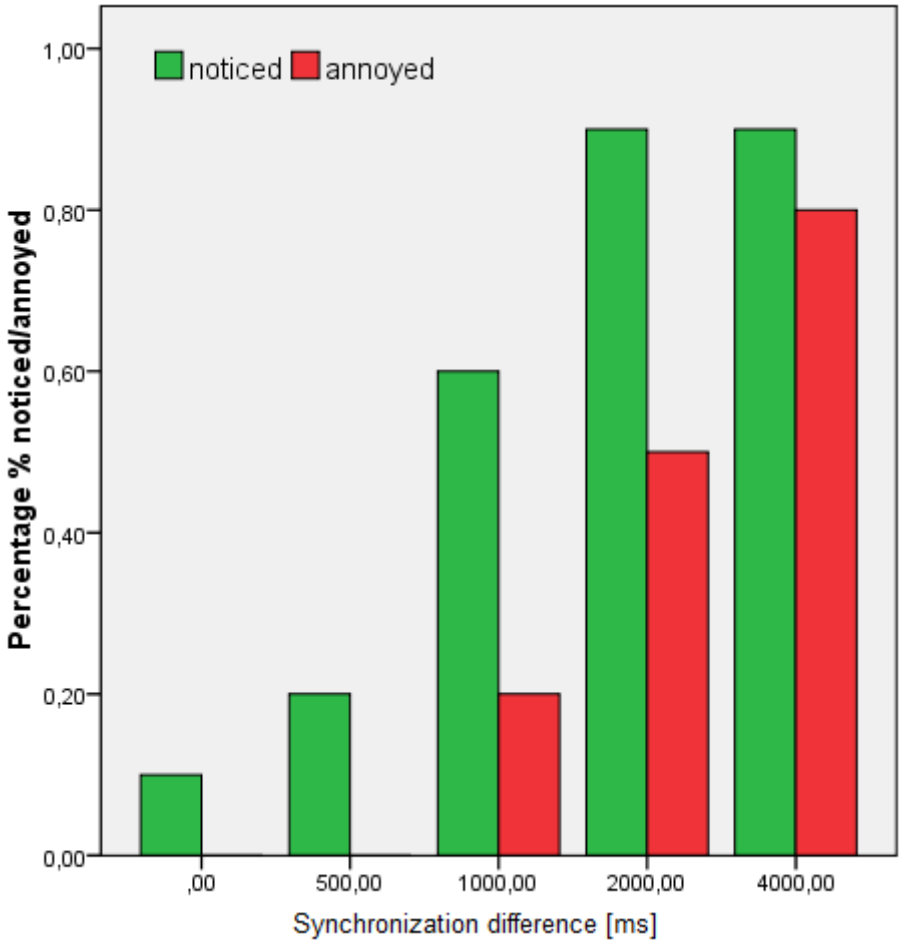


Figure 7.5 Annoyance and perception in soccer watching experiment

The results in table 7.1 show how the results from 7.6 may have an effect on consumer behavior. In the case of 4s synchronization difference, 40% of the participants indicate they would like to switch from their TV provider with similar costs.

Synchronization diff [ms]	NO	MAYBE	YES
0	80%	20%	0%
500	80%	20%	0%
1000	70%	30%	0%
2000	30%	70%	0%
4000	20%	40%	40%

Table 7.1 Percentage of participants willing to change provider because of inter destination synchronization

7.5 Conclusion

This small scale experiment showed that play-out difference when watching soccer is clearly noticed and perceived as annoying. Broadcasting companies should take these effects into account as for large differences more than 40% considers changing provider. While the topic needs further research the results clearly indicate that the play/out delay of soccer matches constitutes a quality aspect related to the broadcast service.

8 Conclusion/Future Work

8.1 Conclusions

1. According to measured play-out differences and the user experience experiments inter-destination media synchronization is most useful when applied to live TV web-streams. The most useful application to TV broadcasts such as DVB, analogue cable and IPTV would be synchronizing between users in different networks/technologies as this is where the largest differences occur.

2. The thresholds when play-out difference becomes annoying or noticable were approximately the same for both the social TV and soccer watching experience. For social TV and soccer with an audio link a difference of 1 second becomes noticeable. The threshold on annoyance was 4s for social TV. TV service companies must take inter-destination synchronization for soccer fans into account as 40% indicated to be willing to change their service because of 4s play-out difference. Inter-destination media synchronization can be considered useful to enhance the soccer watching experience as such play-out differences have been shown to occur in practice.

3. Contrary to the current state of the art in Social TV, inter-destination media synchronization was found in a large scale user experiment to not have an effect on the social experience in a sociable genre. The idea that inter-destination media synchronization enhances the social experience has also been assumed outside academia. Many companies such as clipsync [39] youtubesocial [38] BBC sync [37], Yahoo! Zync [12] market their product on the assumption that inter-destination media synchronization enhances the social experience. In this experiment relationship between the participants, if voice or chat is used and the level of chat activity were found to be factors determining the amount of togetherness.

4. A pilot measurement study in a small geographic area covering three cities in the Netherlands showed that play-out difference ranges from 0 to 5s and is approximately fixed if the channel and the technology/company is known. This is useful for interactive services around broadcast TV as they can use simple offset schemes to obtain rough synchronization. Also it makes the developed measurement system well applicable as it assumes approximately fixed play-out differences.

5. The usefulness of applying inter-destination media synchronization in a single network as often common in video conferencing, has not been shown particularly useful for TV broadcasts in this thesis. For researchers and protocol designers this is an important aspect to take into account as inter-destination media synchronization techniques have often been applied for video conferencing. It would be a mistake to simply transport the methods and assumptions from video conferencing systems. The differences in applying synchronization in video conferencing and TV broadcasts are shown in table 8.1. The range for good QoE in video conferencing is taken from studies [9] and [8], the play-out differences encountered without synchronization from [5] and [4]. Synchronizing between different networks is practically very different to synchronizing within a single network. The exact implications of this difference are outside the scope of this thesis. The amount of synchronization needed for a seamless experience is also very different, in general less accuracy is required for TV synchronization. This factor should make developing synchronization algorithms for TV less difficult compared to video conferencing (not taking into account the network aspects). The typical differences encountered between TV broadcasts compared to video conferencing multicasts are also much larger in the case of TV broadcasts.

	Video Conferencing	TV Broadcasts
Network requiring synchronization	Same network/protocol	Different Networks/protocols
seamless experience	0-200ms range	0-1s range
Typical differences encountered	0-1 s range	0-6 s range

Table 8.1 Differences of inter-destination media synchronization between video conferencing and TV broadcasts/social TV

8.2 Future work

1. More data on inter destination media synchronization should be obtained for different countries and geographical areas. The United States is of particular interest due to its size and its media landscape. In the United States broadcasts are much less centralized as in the Netherlands with local station for example leasing content from larger stations but inserting commercials in a different ways. These particular non technical causes of play-out difference have not been studied in this thesis and need further research

2. An interesting use case for Inter-destination media synchronization is introduced by second screen synchronization. Community gaming around TV content on a second screen poses different synchronization requirements deserve further investigation. Comparison with [9] shows that smaller synchronization bounds might be needed compared to Social TV and soccer watching. This has many applications such as for example rating systems for talent shows and live interactive quiz shows.

3. More specific user tests focusing on specific applications or target groups such as children, elderly and genres should be performed. The current performed test only takes the broad case of a general audience and sociable genre into account. Requirements might be different for specific groups of people.

4. Implementation of specific synchronization solutions that fulfill the recommendations found in this Thesis. These can be either on set-top box, handheld or PC devices.

References

1. **Erik Boertjes, Sven Schultz, Junte Klok.** *Pilot ConnecTV gebruikersonderzoek.* sl : TNO(2008), 2008.
2. **ETSI TS.** *NGN integrated IPTV subsystem stage 3 specification.* sl : ETSI TISPAN, 2008.
3. **Toshiro Nunome, Yutaka Ishibashi.** *Inter–Destination Synchronization Schemes for continuous media multicating, an application level QoS comparison in Hierarchical Networks.* sl : IEICE TRANS. COMMUN., VOL.E85–B, NO.1 JANUARY 2002, 2002.
4. **Toshiro Nunome, Shuji Tasaka.** *Application level QoS comparison of inter-destination media synchronization schemes for continuous media multicasting.* sl : IEICE Transactions on communications vol. 87, 2004.
5. **ITU-T G.1050.** *Network model for evaluating multimedia transmission performance over internet protocol.* sl : ITU-T, 2007.
6. **Reimers, Ulrich.** *DVB- The family of International Standards for Digital Video Broadcasting.* sl : Proceedings of the IEEE 2006, Vol. 9 nr .1, 2005.
7. **Steinmetz, Ralph.** *Human perception of Jitter and Media synchronization.* sl : IEE journal on Selected Areas in Communications Vol 14. issue 1, 1996. 1.
8. **Kazuki Hosoya, Yutaka Ishibashi, Shinji Sugawara.** *Group synchronization control considering difference of conversation roles.* Kyoto : Consumer Electronics, 2009. ISCE '09. IEEE 13th International Symposium, 2009. pp. 948-952.
9. **Yutaku Ishibashi, Manuba Nagasaka, Noriyuki Fujiyoshi.** *Subjective Assesment of Fairness among users in multipoint communications.* sl : Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology, 2006.
10. **David Geerts, Dirk de Grooff.** *Supporting the social uses of television: sociability heuristics for social TV.* Boston : CHI '09 Proceedings of the 27th international conference on Human factors in computing systems, 2009.
11. **Lora Oehlberg, Nicolas Duchenaut, James D. Thornton.** *Social TV: Designing for distributed, sociable television viewing.* Athens : Euro ITV 2006, 2006.
12. **David A. Shamma, Marcello Bastea-Forte, Niels Joubert, Yiming Liu.** *Enhancing online personal connections through synchronized sharing of online video.* Florence : ACM CHI EA '08 extended abstracts on Human factors in computing systems, 2008.
13. **Schulzrinne, Casner, Frederick,Jacobson.** *RFC 3550 RTP: A transport protocol for real-time applications.* sl : ietf, 2003.
14. **Rosenberg, Schulzrinne, Camarillo, Johnston, Peterson, Sparks.** *RFC 3261 SIP: Session initiation protocol.* sl : ietf, 2002.
15. **Hans Stokking, Oskar Van Deventer, Ray van Brandenburg, Omar Niamut, Ishan Vaishnavi.** *RTCP XR Block type for inter-destination media synchronization.* Beijing : ietf, 2010.
16. **Block, Gerry.** HDTV-Gaming la: an epidemic exposed. *IGN Entertainment games.* [Online] 9 3, 2011. [Cited: 9 3, 2011.] http://uk.gear.ign.com/articles/712/712352p2.html.
17. —. The HDTV gaming lag report. *IGN Entertainmanent games.* [Online] 3 9, 2011. [Cited: 3 9, 2011.] http://uk.gear.ign.com/articles/720/720303.
18. **Arogan.** LCD HDTV Input lag tests. *arogan gaming blog.* [Online] 9 8, 2008. [Cited: 3 9, 2011.] http://blog.arogan.com/2008/09/lcd-hdtv-input-lag-tests.html.
19. **Author, Unknown.** HDTV Lag the unofficial guide. *sites.google.com.* [Online] [Cited: 9 3, 2011.] http://sites.google.com/site/hdtvlag/home.
20. **forum, AV.** LCD input lag/video delay measurement (PC and console games responsiveness). *AVS forum.* [Online] 27 8 2007-2008. [Citaat van: 9 3 2007.] http://www.avforums.com/forums/lcd-led-lcd-tvs/612503-lcd-input-lag-video-delay-measurement-pc-console-games-responsiveness.html.
21. **Chung-Lin Huang, Bing-Yao Liao.** *A robust scene change detection method for video segmentation.* sl : IEEE conference on video circuits and systems, 2001. p. vol. 11.
22. **Canny, John.** *A Computational approach to edge detection.* sl : IEEE trans. Pattern Analysis and Machine Intelligence, 1986.
23. **Ginkel, Van.** A matlab toolbox for scientific image processing and analysis. *www.diplib.org.* 2 2009.

24. **Ishfaq Ahmad, Xiaohui Wei, Yu Sun, Ya Qin Zhang.** *Video Transcoding: An overview of various techniques and research Issues.* sl : IEEE, 2005.
25. Digital Video Discussion. [Online] [Cited: 3 9, 2011.] www.doom10.org.
26. **Jacobsen, Eric.** *Metrics and techniques for Evaluating FEC systems.* sl : IEEE, 2000. IEEE 802.16.1pc-00/25.
27. **Publiek omroep Bouquet.** *Radio en TV Zenders in Nederland.* [Online] 2011. http://www.radio-tv-nederland.nl/dvbt/digitenne-kpntv.html.
28. **Phaeron.** Virtual Dub. [Online] http://www.virtualdub.org/.
29. **various.** HDTV and Game lag: The problem and the solution. *avsforum.* [Online] 2007-2010. http://www.avsforum.com/avs-vb/showthread.php?t=558125.
30. **NOS.** Nederlandse omroep stichting. *Nederlandse omroep stichting.* [Online] NOS. www.nos.nl.
31. **Yue Lue, Benny Fallica, Fernando A. Kuipers, Robert Kooij, Piet Van Mieghem.** Assesing the Quality of Experience of SopCast. *Assesing the Quality of Experience of SopCast.* International Journal of Protocol Technology : sn, 2009. Vol. 4, 1.
32. **Fernando Boronat, Jaime Lloret, Miguel Garcia.** *Multimedia group and inter-stream synchronization techniques: A comparative study.* sl : Journal on information systems vol. 9, 2009.
33. **ITU-T.** *ITU P.800 Mehods for objective and subjective quality assesment.* sl : ITU, 1993-1996.
34. **Elaine Huang, Gunnar Harboe, Joe Tullio, Ashley Novak, Noel Massey, Crysta Metcalf, Guy Romano.** *Social Television comes home:"a field study of communication choices and prectices in TV-based text and voice chat".* Boston : Proceedings of the 4th Euro TV conference, Athens, Greece (2006) 251-259 11, 2009.
35. **Justin D. Weisz, Sara Kiesler, Hui Zhang, Yuqin Ren, Robert E. Kraut, Joseph A. Konstan.** *Watching Together: Integrating Text Chat with Video.* San Jose : CHI '07 Proceedings of the SIGCHI conference on Human factors in computing systems, 2007.
36. **Geerts, David.** *Comparing voice chat and text chat in a communication tool for interactive television.* Oslo : sn, 2006.
37. **Rose, Anthony.** BBC Internet blog. *Introducing the all new I-player.* [Online] [Cited: 3 9, 2011.] http://www.bbc.co.uk/blog/bbcinternet/2010/05/introducing_the_all_new_bbc_ip.htmlThe 13th IEEE International symposium on Consumer Electronics The 13th IEEE International symposium on Consumer Electronics.
38. **youtubesocial.** youtubesocial. [Online] http://youtubesocial.com.
39. **Clipsync.** www.clipsync.com. [Online] 9 3 2011. clipsync.com.
40. **David Geerts, Pablo Cesar, Dick Bulterman.** *The implications of program genres for the design of social television systems.* San Francisco : ACM, 2008.
41. **Brinkman, Willem Paul.** *Slides IN4303 Emperical research methods.* Delt : sn, 2009.
42. **Tara Shrimpton-Smith, Bieke Zaman, David Geerts.** *Coupling the users: The benefits of paired user testing for iDTV.* Athens : Ablex Pub. Corp. International Journal of Human-computer Interaction vol:24 issue:2 pages:197-213, 2006. pp. 214-211.
43. **Jurgen Vogel, Volker Hilt, Wolfgang Effelsberg.** *W. Local-lag and timewarp: Providing consistency for replicated continuous applications.* Los Angeles : IEEE, 2004.
44. teamspeak communication system. [Online] 3 9, 2011. [Cited: 3 9, 2011.] www.teamspeak.com.
45. **Gustav Theodor Fechner, Hartel Breitkopf.** *elemente der Psychophysik.* Amsterdam : Bonset, 1860/1966.
46. **NIST.** NIST/SEMANTECH. *Engineering statistics handbook.* [Online] 2010. [Cited: 9 3, 2011.] http://www.itl.nist.gov/div898/handbook/.
47. **Oskar van Deventer, Hans Stokking, Omar Niamut, Fabian Walraven, Rufael Mekuria.** Inter destination. *Inter destination Media Synchronization, now standardized by ETSI TISPAN.* [Online] 2010. [Cited: 3 9, 2011.]
http://www.ngnlab.eu/main/images/2nd_ngnlab_eu_workshop/tno__etsi_synchronization.pdf.
48. **ITU-T.** *Interactive test methods for audio-visual communications* . sl : ITU-T, 1996.
49. Broadcast Magazine Netherlands. 2010. Vol. 8.
50. **NWO.** Nationale wetenschaps quiz. *www.nwo.nl/quiz.* [Online] nwo, nederlandse organisatie voor wetenschappelijk onderzoek. www.vpro.nl.

51. **Margeret Pinson, Patrick Wolf.** *Reduced refererence method using causality processing for estimating variable video delays.* sl : NTIA, 2007.
52. **ETSI TS.** *ETSI TS 182027 IPTV architecture IPTV functions supported by the IMS subsystem.* sl : ETSI TISPAN, 2008.
53. **Benny Fallica, Yue Lu, Fernando Kuipers, Robert Kooij, Piet Van Mieghem.** *On the quality of experience of SopCast.* Cardiff : IEEE Next generation mobile applications conference, 2008.
54. **David Geerts, Ishan Vaishnavi, Rufael Mekuria, Oskar Van Deventer, Pablo Cesar.** *Are We in sync? Synchronization requirements for watching online video together.* Vancouver : ACM CHI'11, 2011.
55. **Robert Kooij, Kamal Ahmed, Kjell Brunnstrom.** *Perceived Quality of channel zapping.* Palma de Mallorca : Fifth IASTED Conference, 2006, 2006. pp. 155-158.
56. **Fernando Kuipers, Robert Kooij, Danny De Vleeschauwer.** *Techniques for measuring quality of experience.* Lulea : WIRED/WIRELESS INTERNET COMMUNICATIONS Volume 6074/2010, 216-227, 2010. pp. 216-227.
57. **Rufael Mekuria, Hans Stokking, Oskar van Deventer.** *Automatic Measurement of Play-out Differences for Social TV, Interactive TV, Gaming and Inter-destination media synchronization.* Delft : Submitted to the EuroITV 2011, 2011.
58. **Raimund Schatz, Siegfried Wagner, Sebastian Egger, Norbert Jordan.** *Mobile TV becomes social, Integrating content with communications.* CavTat : IEEE 29th International Conference on Information Technology Interfaces, 2007. ITI 2007., 2007. pp. 263-270.
59. **Robson, Colin.** *Real world research.* sl : Wiley and Sons, 2002.
60. **Hans Stokking, Oskar van Deventer, Omar Niamut, Fabian Walraven, Rufael Mekuria.** *IPTV Inter-destination synchronization: A network based approach.* Berlin : 2010 14th International Conference on Intelligence in Next Generation Networks (ICIN), 2010.
61. **Ehrenstein, Walter. H.** *Modern Techniques in Neuroscience research.* sl : Springer verlag, 1999.
62. **Wolf, Patrick.** *A full reference method using causality processing for estimating variable video delays.* sl : NTIA, 2009.
63. **Brinkman, Willem Paul.** *Handbook of Mobile Technology Research Methods.* Delft : Nova science methods, 2009.
64. **ETSI TS.** *ETSI TS 181016 "service layer to integrate NGN services and IPTV".* sl : ETSI TISPAN, 2008.
65. **Tumcat.** Testbed for user experience of mobile aware applications. [Online] http://www.tumcat.nl.
66. **X264.** X264. [Online] [Cited: 3 9, 2011.] www.X264.com.

Appendix

A. Structure and manual of the code of the measurement tool

The synchronization measurement demo is started by running idms_detector_demo,m which runs the three scripts shown in table A.1. The parameters for configuring the detector are shown in Table A.2 while the different specific routines are explained in table A.3.

Video_segmentation.m	Segments video to left and right directories
Obtain_scene_data.m	Obtains scene data from the segments
Play-out difference	Computes differences in play-out from obtained scene data

Table A.1 Files for the idms_detector_demo

USER_SELECT	1= user selects file 0=preset file
VIDEO_FILE	The location of the videofile if USER_SELECT=0
V_START	The starting position of the measurement in frames
V_STEP	The step/segment size of the video in blocks
NR_SEGMENTS	The maximum number of segments to use in the computation

Table A.2 Parameters for the parameter_config.m file for the detector demo

Routine	function
dc_diff.m	Computes DC differences between frames
Dc_val.m	Computes DC value frames
Getgray.m	Converts video array to gray scale
Hist_diff	Computes histogram differences
Peak_ratios.m	Computes ratios between peaks in a sliding window
Sort_scenes	Orders results of scene changes obtained
Static_edge_test.m	Performs the static edge test
Static_scenes.m	Computes static scene-change detection
Sync_vek	Computes the scene change sequence function
Syncdiff.m	Computes synchronization difference

Table A.3 The files of the scene detection mechanism

B. Structure and manual of the synchronization system from chapter 7

VLC_run3.jar	Run at the administrator to control the two terminals under control. The arguments are [synchronization difference ms] [ip address 1] [ip address 2]
Play.html	Vlc script that controls the media player at the host. It starts up a new video

Table B.1 Files for the soccer video synchronization mechanism

C. DVD Contents

Datasets	Contains datafiles for SPSS/excel for soccer and Leuven experiments
Images	Contains images from this thesis and plots not included
Measurements	Contains video measurement data of broadcast TV and web streams
Presentations	Contains intermediate and final presentations of this work
Publications	Contains publications to which the author contributed, published or submitted to conferences
Software	Contains the software for the measurement system and the synchronization mechanism for the TV experiment

Table C.1 Contents of the directories in the supplementary DVD

D. (Co)-authored publications

Hans Stokking, Oskar van Deventer, Omar Niamut, Fabian Walraven, Rufael Mekuria. *IPTV Inter-destination synchronization: A network based approach.* Berlin : 2010 14th International Conference on Intelligence in Next Generation Networks (ICIN), 2010.

Rufael Mekuria, Hans Stokking, Oskar van Deventer. *Automatic Measurement of Play-out Differences for Social TV, Interactive TV, Gaming and Inter-destination media synchronization.* Delft : Submitted to the EuroITV 2011, 2011.

David Geerts, Ishan Vaishnavi, Rufael Mekuria, Oskar Van Deventer, Pablo Cesar. *Are We in sync? Synchronization requirements for watching online video together.* Vancouver : ACM CHI'11, 2011.

E. Demo of measurement system Software

The pictures below show the system in practice. The file idms_detector_demo is run from the MATLAB environment. The .wmv file with the recording is selected from the dialog box as shown in Figure E.1. The user segments the two different screens as shown in Figure 2.13 subsequently the video file with the recording is segmented in left and right frames to be processed by the scene-change detector as shown in Figure E.2. After this the scene changes are computed as shown in E.3. After these steps the synchronization result is given at the command line as output Figure E.4.

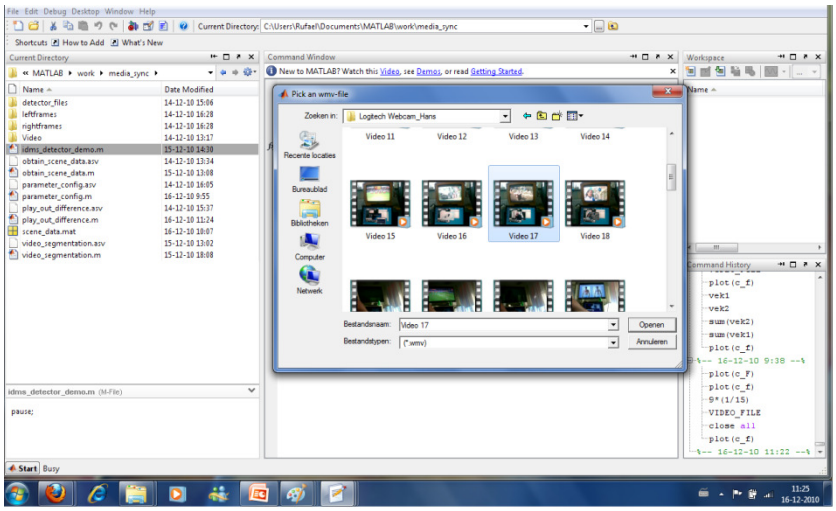


Figure E.1 Measurement system: recorded file selection

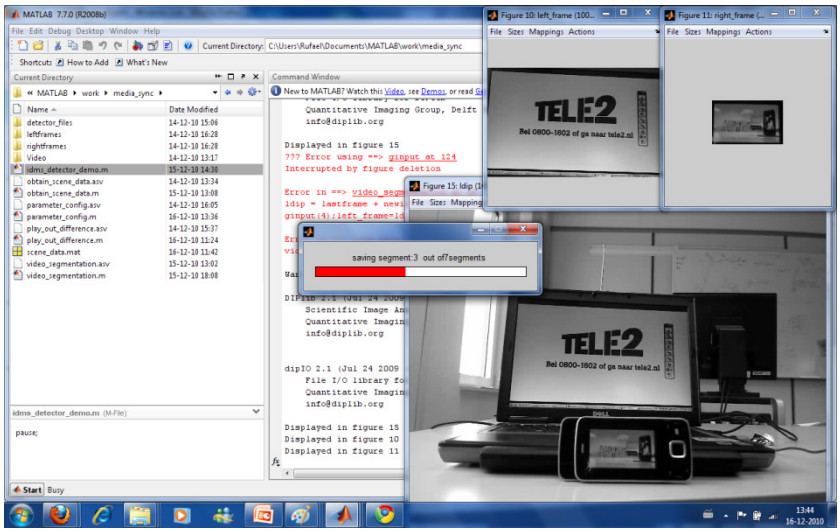


Figure E.2 Measurement system: segment extraction

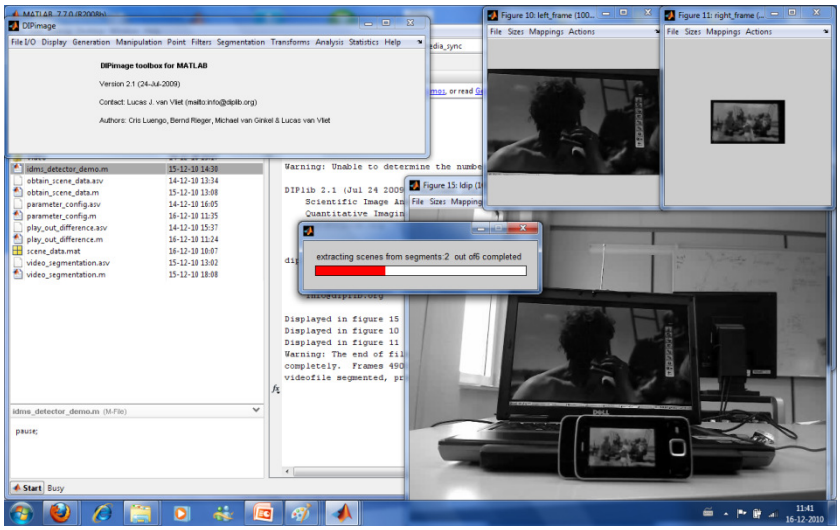


Figure E.3 Measurement system: scene change computation

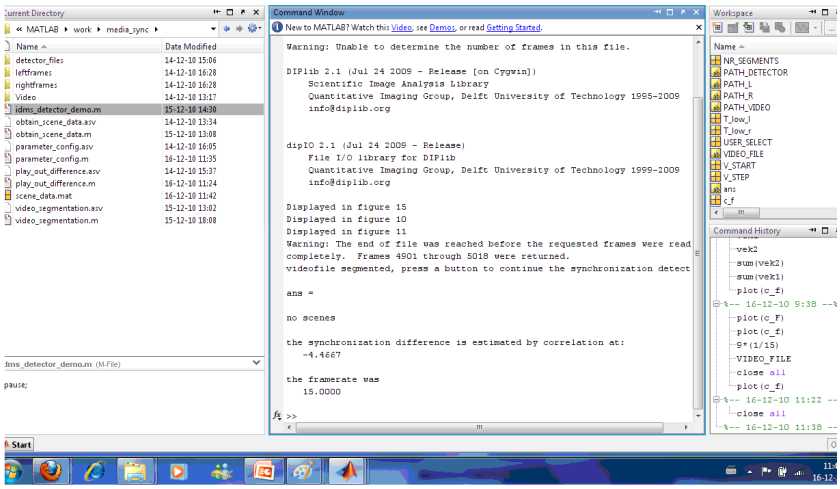


Figure E.4 Final result displayed on the command line