Are We in Sync? Synchronization Requirements for Watching Online Video Together

David Geerts*, Ishan Vaishnavi†, Rufael Mekuria‡, Oskar van Deventer‡, Pablo Cesar†
*CUO, IBBT / K.U.Leuven, †CWI, ‡TNO
david.geerts@soc.kuleuven.be, oskar.vandeventer@tno.nl, P.S.Cesar@cwi.nl

ABSTRACT
Synchronization between locations is an important factor for enabling remote shared experiences. Still, experimental data on what is the acceptable synchronization level is scarce. This paper discusses the synchronization requirements for watching online videos together – a popular set of services that recreate the shared experience of watching TV together by offering tools to communicate while watching. It studies the noticeability and annoyance of synchronization differences of the video being watched, as well as the impact on users' feelings of togetherness, both for voice chat and text chat. Results of an experiment with 36 participants show that when using voice chat, users notice synchronization differences sooner, are more annoyed and feel more together than when using text chat. However, users with high text chat activity notice synchronization differences similar to participants using voice chat.

Author Keywords
Online video, Synchronization, Social TV, Entertainment

ACM Classification Keywords
H.4.3 Communications Applications, H.5.1 Multimedia Information Systems

General Terms
Experimentation, Measurement

INTRODUCTION

Traditionally, watching television is for a large part a social activity as viewers in the same location often discuss the contents of the programs they are jointly watching [5]. Recently, online video sites like ClipSync and Watchitoo try to recreate this shared experience over the Internet by adding text chat features or an audio channel alongside the videos on offer. This allows users to watch a synchronized version of a video while communicating with each other. These examples are part of a larger trend of integrating social media and communication features with video content, not only on the Internet, but also on traditional television sets [2] and even mobile phones [7].

An important assumption when offering users the option to communicate while watching video together is that the video needs to be synchronized in order to have some common ground to talk about [9]. While viewers do not always talk about the video content while watching, synchronization is important when the content of the program is the topic of the conversation. However, it is theoretically impossible to exactly synchronize video play-out over a network. Therefore a key research question is to know which synchronization level still enables users to have a satisfying shared experience, impacting the design of future social video systems.

This paper examines this issue by discussing the results of an experimental user study focusing on synchronization differences of videos when jointly watched video content while communicating using either voice chat or text chat.

RELATED WORK

Previous research on social video watching has concentrated on studying communication choices based on results of field trials [2] and on identifying appropriate sociability heuristics [1]. Other relevant work had a more specific focus, e.g. on measuring the level of togetherness between friends or strangers [9] or on identifying in detail user activity while watching together [8]. The findings of all this work points to a common direction: a direct communication link between people watching video together is desirable and it increases the level of togetherness.

Apart from the general conclusion, these and other studies reveal specific issues which can influence the shared video watching experience. Some of the parameters that might affect communication while watching videos include how well people know each other, which genre they are watching [1], if they like the video they are watching and what communication modality they use [2]. Additional factors that may play a role as well are if users have seen the video before, what is happening in the program at a specific moment, or even a person's personal characteristics.

On top of these results, a particular question that has not been tackled in the past is what the acceptable difference is in synchronization while watching television together. Currently, 150ms is used as a rule of thumb, a value drawn from telecommunications research. This rule states that the

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

maximum end-to-end, one-way delay when talking remotely should not be over 150ms [4]. Below this value users cannot perceive the delay in communication, and therefore cannot detect differences in synchronization of shared video content. However, no actual user studies have been done to determine the range of acceptable synchronization levels for social video watching. This, we believe, is in part because of the number of parameters that this value may depend on, as described above.

This work intends to answer part of this question by presenting the results of a user study, which isolates some of these parameters and determines the relevant acceptable synchronization levels for those parameters, as well as their impact on users' feeling of togetherness.

METHODOLOGY

Test Setup

A within-subjects lab-based experiment was conducted with 18 couples (partners, friends, or family), with a total of 36 people taking part in the tests, consisting of 12 males and 24 females. The age ranged from 15 to 68 years old, which is wider spread than in most previous research, reflecting the broad target audience of shared video watching.

Each participant from each couple was shown two episodes of a popular local quiz show at different locations. The show was chosen because a quiz is a very sociable genre [1], and it was carefully edited to offer consistent content during the test. During the first episode, participants could voice chat with each other using a headset. The headset was also used for listening to the audio track of the video content, so the participants could not hear the audio track of their partner's show. During the second episode, participants could only text chat with each other. The text chat was implemented with a chat box on the same screen as the video (on a laptop), positioned at the right side of the video (as in services like YoutubeSocial or Watchitoo). Messages were sent line-by-line, as is common in most chat services. The tests were carried out on a private LAN with no external influence presumably limiting the end-to-end voice/text chat delay. We did not observe any noticeable delay in communication between users. The order of text chat and voice chat conditions was randomized over the different test sessions, in order to remove any habituation effects.

Without informing the participants, every seven minutes the synchronization level of the videos was changed. This length was chosen in order to allow participants enough time for having a substantial conversation, as well as being able to present several conditions to participants during a two-hour test session. In each condition (voice chat and text chat), five synchronization levels were presented to users: 0 seconds (perfect sync), 500 milliseconds, 1 second, 2 seconds and 4 seconds. These values were chosen during a test by two of the authors, in which it was discovered that video synchronization difference becomes detectable between 500ms and 2s. Based on these results - and supported by an earlier pilot study with 21 users - 500ms, 1s, and 2s were taken as test condition, and 0s and 4s were chosen to test the more extreme cases. These levels were presented in a randomized order for each set of participants and each condition. As a difference in synchronization between two participants implies that (the video of) one person is ahead ('leading'), and one person is behind ('lagging'), the order of who is leading and who is lagging was also randomly varied.

After each seven minutes (before the next synchronization change), the participants were asked to fill in a web-based questionnaire, asking a series of questions related to togetherness, noticeability and annoyance of the synchronization differences. For measuring togetherness, six questions were asked (e.g. "I felt 'together' with my partner"). These six questions were tested to be consistent (Cronbachs alpha α=0.852, and Gutmann's split half 0.807). From these six questions an aggregate measure was derived (the average) to indicate the togetherness experienced by the participant. To measure the noticeability and annoyance of synchronization differences, the Degradation Category Rating (DCR) MOS score as described in [3], used for degraded speech signals, was adapted with values ranging from 1 (not noticeable) to 5 (noticeable and very annoying).

In total each participant filled in 10 questionnaires (5 during each condition), resulting in 360 unique measurements. After the first questionnaire it would become clear that synchronization was one of the issues which was questioned. Therefore the participants were instructed in advance to only talk about the content of the show, and not discuss the test itself nor explicitly try to figure out the synchronization difference of the videos.

Synchronization Algorithm

In order to control the synchronization during the user tests, a system is required which can play media synchronized in two locations and can be manipulated by the observers. A simplified version of the local lag algorithm [6] was used to achieve the chosen level of synchronization. One of the participants' computers was chosen as a master, which continuously sent out position updates to the other computer (the slave). The slave computer received these updates and jumped to the recommended position. Before the tests were conducted, this mechanism was validated within the test environment and a margin of error was established for the synchronization levels. It was found that the error in synchronization levels was maximum 150ms with an average of 8ms difference and a standard deviation of 59ms. Thus in this experiment a synchronization level of 0 implies an interval of 0 +- 0.15 seconds.

Hypotheses

As the two conditions being tested were talking (voice chat) and chatting (text chat), the analysis will mainly focus on the differences between both modalities. Therefore the following three general hypotheses are formulated:

H1 People feel more together when using voice chat than when using text chat

H2 People will notice synchronization difference sooner when using voice chat than when using text chat

H3 People will be more annoyed by synchronization differences when using voice chat than when using text chat

For each of the hypotheses, the influence of other factors was tested such as chat experience, chat activity, or if the participants liked the program or not. For testing text chat activity, the participants were divided into an active group (more than 400 words per session, which was close to the median), with N=15 participants, and a non-active group (less than 400 words per session), with N=21 participants.

Play-out differences and the use of text/voice chat were taken as explanatory variables. The dependent variables measured in each condition on each participant were noticeability, annoyance and togetherness. Repeated measures analysis was used to calculate the effect taking within subject effects into account. Interaction effects were analyzed with a between-within analysis.

RESULTS

Togetherness

The answers on the togetherness questions show that voice chatters feel more together on average than text chatters (one way ANOVA $F(1,140)=14.26$, $p<0.01$). Although significant, the difference is small as on average it was approximately one point on a 7 point scale (ranging from 1, completely not together, to 7, completely together). This corresponds to text chatters being "somewhat together" (mean togetherness=4.3) on average and voice chatters being mostly "together" (5.1) on average. Although we expected to find different scores depending on how well the video was synchronized, the effect of synchronization levels on togetherness was not found significant.

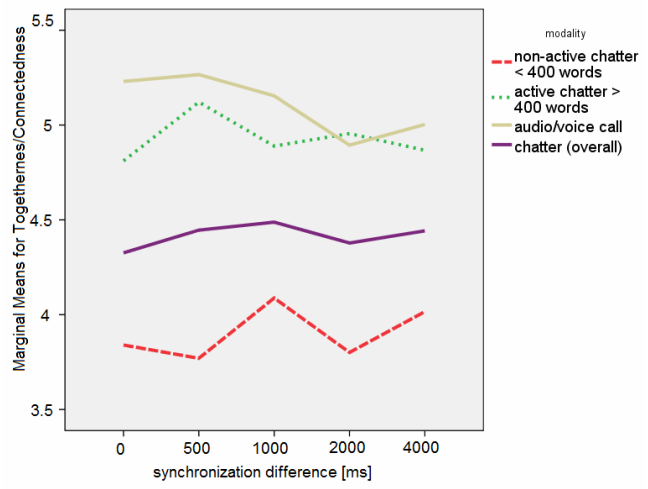

Figure 1: togetherness by chat activity

When taking into account chat activity (see Figure 1), this difference in mean togetherness between voice chatters and text chatters is mainly attributed to non-active chatters (3.9). Between active text chatters (4.9) and voice chatters no significant difference on togetherness was found. Taking these three groups into account, voice chatters and active chatters feel significantly more together than non-active chatters ($F(1,359)=93.5$, $p<0.001$). This means that while H1 in general might be true, it has to be rejected when comparing active text chatters with voice chatters.

Noticeability and Annoyance

H2 focuses on the noticeability of synchronization, while H3 concerns annoyance. As both are closely related, they will be discussed together. Figure 2 shows that in the voice chat condition, synchronization difference becomes noticeable for most people when it is 1s or more.

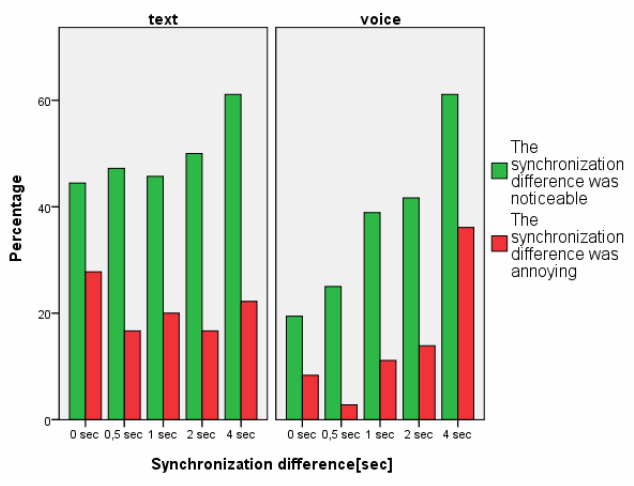

Figure 2: Noticeability and annoyance of play-out differences

People in the text chat condition give rather random answers, not correlated with the synchronization levels, indicating that they do not notice a difference based on synchronization level, but probably attribute this to other factors (such as reaction time of the other participants).

The statistical results show that synchronization differences in the 0-4s range tested were noticed significantly by voice chatters ($F(4,140)=6.479$, $p<0.001$). The effect of synchronization difference on annoyance was also found significant for the voice chat case ($F(2.33, 81.66) =6.845$, $p<0.05$). Text chatters however, did not notice synchronization differences significantly more or less often for each different level ($F(4,140)=0.887$, $p>0.05$). Also the effect of the synchronization difference on annoyance was not verified ($F(4,132)=0.564$, $p>0.05$). Based on these results, H2 and H3 can be accepted, as voice chatters notice synchronization more easily, and are more annoyed by it.

It is interesting to see if the likeability of the content, or the fact that the participants had seen the episodes, would influence the noticeability of synchronization differences.

Overall, voice and text chat participants that liked the show seemed to notice synchronization difference more quickly than people that were neutral to it. This effect however was tested non-significant ($F(8,132) = 0.859$, $p>0.05$). Having seen the episode before also does not make synchronization difference more or less noticeable ($F(1,359)=0.875$, $p>0.05$).

Text chat experience and text chat activity were also tested as mediating factors. Experienced text chatters (more than once per month) do not notice synchronization difference better than less experienced text chatters (less than once per month) ($F(4,112) = 0.029$, $p>0,05$).

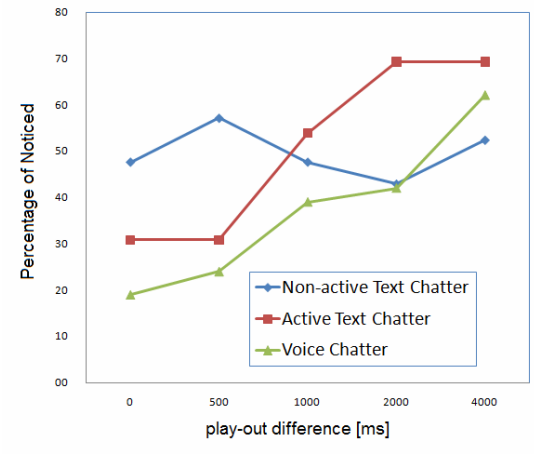

Figure 3: Noticeability by chat activity

The difference between active chatters and non-active chatters however was found significant ($F(1,32)=6.116$, $p<0.05$). Figure 3 shows that active text chatters are able to notice synchronization differences larger than 1 second, similar to voice chatters. Due to the few participants that got annoyed, no similar claims on annoyance can be made.

CONCLUSION

This paper discussed how well people using voice chat and text chat notice synchronization differences when watching online videos together, and what the impact is on annoyance and togetherness. While currently telecom operators are aiming at the synchronization level found in telecommunication tests (150ms) our results show that voice chatters only start noticing differences above 2 seconds delays. Most text chatters do not notice synchronization differences between 0 and 4 seconds, however active text chatters notice synchronization differences similar to when using voice chat. As the highest levels of togetherness were also observed with active text chatters and all voice chatters, we recommend synchronization of approximately 1 second (which was not noticeable by this group) for a seamless shared experience. These results put into doubt the 150ms value from telecommunications research as the target synchronization bound required for social video watching applications. A first implication for software designers is that they can concentrate on implementing simpler mechanisms that aim at a synchronization level of 1second. Even more interesting, these results imply that social video watching applications can be effective even on more challenging platforms such as mobile networks, where delays are unpredictable, or thin clients, where execution times may be variable. This will result in new opportunities for network operators and system designers, which in turn may provide more flexibility and dynamism to end-users in future ubiquitous social video applications. Further research should focus on the influence of other parameters such as different genres or different platforms (PC, mobile, TV) to test whether these results are also valid in other circumstances.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2007-214793

REFERENCES

[1] Geerts, D. and De Grooff, D. 2009. Supporting the social uses of television: sociability heuristics for social TV. In Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 595-604.

[2] Huang, E. M., Harboe, G., Tullio, J., Novak, A., Massey, N., Metcalf, C. J., and Romano, G. 2009. Of social television comes home: a field study of communication choices and practices in tv-based text and voice chat. In Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 585-594.

[3] ITU-T Rec. P.800 “Methods for subjective Determination of Transmission quality”, August 1996

[4] ITU-T G.114: "General Recommendations on the transmission quality for an entire international telephone connection, One-way transmission time", May 2003.

[5] Lull, J. 1980. Family Communication Patterns and the Social Uses of Television. In Communication Research, 7(3):319--333.

[6] Mauve, M., Vogel, J., Hilt, V., Effelsberg, W. 2004. Local-lag and timewarp: Providing consistency for replicated continuous applications. IEEE transactions on Multimedia

[7] Schatz, R., Wagner, S., Egger, S., and Jordan, N. 2007. Mobile TV becomes social: integrating content with communications. In Proceedings of the ITI Conference on Information Technology Interfaces, pp. 263-270.

[8] Shamma, D. A., Bastea-Forte, M., Joubert, N., and Liu, Y. 2008. Enhancing online personal connections through the synchronized sharing of online video. In CHI '08 Extended Abstracts on Human Factors in Computing Systems, pp. 2931-2936.

[9] Weisz, J. D., Kiesler, S., Zhang, H., Ren, Y., Kraut, R. E., and Konstan, J. A. 2007. Watching together: integrating text chat with video. In Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 877-886.