On Time or Not on Time: A User Study on Delays in a Synchronised Companion-Screen Experience

Christoph Ziegler[1], **Christian Keimel**[1], **Rajiv Ramdhany**[2] **and Vinoba Vinayagamoorthy**[2]

[1]IRT, Munich, Germany, ziegler@irt.de
[2]BBC R&D, London, United Kingdom, vinoba.vinayagamoorthy@bbc.co.uk

ABSTRACT

One major challenge in creation of compelling companion screen experiences, are the time delays between the presentation of content on the TV compared to the presentation of content on the companion screen. Through the use of a synchronised, interactive textbook application, we conducted a user study to evaluate the potential influence of different delays, between the TV and the companion screen, on how users experience watching a Shakespearean play on the TV. Our results indicate that although users do not notice delays of up to 1000 ms, for the kind of experience tested, they feel significantly more distracted by the tablet content for increasingly higher delays. We discuss the implications of our findings with regards to the time delay tolerances users might have when using a synchronised text accompaniment to these kinds of TV programmes.

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces – User-centered design

Author Keywords

Companion screen; connected experiences; interaction techniques; second screen; television; synchronisation; impact of delays.

INTRODUCTION

The ubiquity of smart devices, such as mobile phones and tablets, and their widespread use in almost every aspect of human life including, TV viewing in the home, is well documented in market research surveys such as [14, 24].

Previous studies have stressed the untapped potential of additional services that accompany the TV programme on tablets or smart phones. This includes engaging viewers further with the programme and thus enriching the overall TV experience [25]. Although the concept of presenting accompanying content on companion devices is not entirely new, a key component for creating compelling *companion applications* is the technical ability to enable tight synchronisation of content presentation on TV and companion devices as demonstrated by [27, 28].

Different solutions for achieving companion screen synchronisation exist. They differ in underlying technical principles, end-users' hardware requirements, delivery mechanisms, and achievable synchronisation accuracy. These are important considerations for broadcasters or content providers in terms of the design decisions and commitments implied by the enabling technology. Whilst perfect zero-delay synchronisation of companion content with the TV remains the ideal case, it may not be cost effective to achieve, or even technically possible by commodity devices. Low-delay sync may not even be necessary for particular types of user experiences; a sample-accurate media synchronisation solution may be superfluous if delays of the order of a few seconds in magnitude are not noticeable by users.

We argue that it is essential to determine how the synchronisation inaccuracy (delay) impacts the user experience[1]. This knowledge will enable broadcasters and content providers to evaluate and optimise the trade-off between user experience and synchronisation accuracy based on their distribution platforms and/or media-synchronisation technologies. A better understanding of the perception of delays and their impact on the user experience of synchronised companion applications would also help in the design of degradation of services offered in companion screen applications based on the technologies being used in homes.

In this study, we examine the impact of two main factors. The first is the delay of presentation of content between the TV and the companion device. The second is three different types of interaction offered to the user in the companion application. This study is based on a potential use case involving the synchronised presentation of the transcript of a Shakespeare play (*Richard II* [4]) in time with a video recording of the play on the TV. This combines the use of subtitles in TV programmes with *surtitles* often used in theatres to increase the accessibility and comprehensibility of performances. The script on the companion screen is also augmented with synchronised highlighting of the text based on the elocution of lines by characters in the scene. The companion application developed for this study further allows for three different types of interaction modes: *passive*, *exploration* and *call-to-action*, allowing us to assess the impact of different types of interactions on the users' experience.

[1] A person's perceptions and responses that result from the use or anticipated use of a product, system or service[15].

TVX '17, June 14-16, 2017, Hilversum, Netherlands

© 2017 ACM. ISBN 978-1-4503-4529-3/17/06…$15.00

DOI: http://dx.doi.org/10.1145/3077548.3077557

In the following sections we present related work in human factors with respect to media synchronisation, we discuss their applicability to companion screen scenarios and examine potential shortcomings. We then provide an overview of the companion screen application developed for this study. Then, we layout a description of the experimental design and state the hypotheses to be addressed by the study presented. Finally, we discuss the results of the experiments and conclude with a short summary of our key findings.

RELATED WORK

Understanding when media streams are perceived to be synchronised is helpful in determining the requirements for the temporal accuracy of different streams within single devices and between multiple devices.

Steinmetz et al. [26] studied the perception of delays between audio and video streams to find the permissible delay at which streams were perceived to be in "*lip sync*". Their studies provided a threshold for absolute delays in addition to an indication that users may adapt to constant delay between audio and video. Murray et al. [21] investigated perception of synchronisation between olfactory data and video. They observed that perception thresholds are significantly different if participants were presented olfactory data and video with audio as opposed to olfactory data and video only. This implies that there are different cues for perception of skews in synchronisation.

Looking at a scenario similar to the type of applications relevant to this paper, there are studies that explored the impact of timing on the perceived quality of TV subtitles [1, 5, 17]. Similar to findings from the study on synchronisation between audio and video, a threshold effect was observed in the studies on subtitle timing. Of course, for subtitles, the users' attention is not split between two devices as it inevitably will be for experiences delivered concurrently between multiple screens as discussed in this study.

Other studies examined media streams played back at different devices at different physical locations, a scenario that is also referred to as inter-destination media synchronisation (IDMS). Montagud et al. [19] reviewed use cases for IDMS. Technical solutions for IDMS have been proposed for example by [2, 16, 20]. Geerts et al. [10] investigated delays between video presentation and simultaneously using text or voice chat, identifying thresholds depending on the frequency of the chat messages and modality. The test subjects, however, were not exposed to two synchronised media streams, but had processed implicit clues from their chat partners in order to detect delays. Mu et al. [20] observed a content dependency of the delay between audio and video streams on different devices on the Qualitity of Experience, suggesting that high temporal complexity content may mask delays better than low temporal complexity content. The content on the different devices, however, was the same, unlike the typical companion screen scenario.

STUDY OBJECTIVES AND HYPOTHESES

Vinayagamoorthy et al. [28] theorised that synchronisation accuracy requirements depend on the type of companion-screen experience: responsive guide app (~1 s) versus a companion application which delivers spatial audio effects (10 μs to 10 ms). There is evidence from previous research which supports this assumption. For example, in a study on the impact of delays on the Quality of Experience during synchronised audio presentation different companion devices, Mu et al. [20] found content genre is a determining factor on the thresholds values for delay perception (noticeable-delay and annoying-delay) by viewers. We argue that content genre is a generalisation of content characteristics (including interaction-complexity) in terms of the content factors that influence the way viewers perceive and experience delays in a companion-screen scenario. Research needs to systematically explore the impact of different dimensions of content characteristics.

We selected *interaction* as one of the dimensions of content-complexity to include in our study. The effect of interaction on delay-perception during a synchronised companion screen experience is of significant interest since interaction is a key element of the user experience in many companion applications (e.g. quiz or poll applications). There is no reported insight in the literature about how different types of interaction affect delay perception in synchronised companion screen experiences.

Further, the focus of previous user studies on media synchronisation has been on the impact of delays on the Quality of Experience (QoE)[2]. We consider the determination of QoE deterioration function based on delay-tolerance thresholds to have been amply covered by previous studies, for example [20, 21, 26]. These studies indicate that delay-tolerance thresholds (level at which delays start getting annoying) are always higher than the perception[3] thresholds (the delay value at which users start noticing delays).

Instead, this study focuses on aspects of the user experience such as visual attention split as indicated through gaze behaviour. It is plausible to theorise that more subtle aspects of the user experience might be affected before users start actively noticing delays. *Attention split* is a salient UX factor as it interferes with the editorial message content creators want to convey in their programme. The goal in the creation of many companion applications is that they should not distract users from the key content delivered on the TV screen, but rather add to the TV experience. If there is a threshold at which delays start driving the user's attention, this threshold can be used as a decision criterion for selecting a specific companion-screen synchronisation system.

Hypotheses

In order to address the question of the influence of delays on the user experience in the proposed application scenario, we formulated the following five hypotheses to be examined in our study:

H1 *At a higher interaction levels, participants have a higher threshold for perception of delays.*

[2] the degree of delight or annoyance of a person experiencing an application, service, or system [18]

[3] the conscious processing of sensory information [18]

H2 *At higher delay levels participants spend more time looking at the companion.*

H3 *At higher delay levels participants' gaze switches more frequently between TV and companion.*

H4 *At higher delay levels participants interact less frequently with the companion application.*

H5 *At higher delay levels participants feel more distracted by the companion content.*

Hypothesis H1 builds on the assumption that interaction requires the users' attention and that users are less sensitive to delays, if their attention is shifted from the synchronisation cues to other aspects of the application.

Hypotheses H2 and H3 build on the assumption that users anticipate a certain timed behaviour from the companion experience. They rely on the companion screen showing a certain piece of content at a certain point in time, without having to look at it. If this expectation is not met due to delays, users might start reading along on the companion or might start looking back and forth between screens to contextualise the presentations on both devices, which results in a higher dwell on tablet or in a higher gaze change frequency.

Straining themselves to make sense of the relation between tablet and TV content, users might feel distracted by the additional screen, which explains hypothesis H5. Furthermore, the overload might hinder them from interacting with the companion application, which justifies hypothesis H4.

THE COMPANION SCREEN EXPERIENCE

The candidate user experience selected for the study involves a TV emulator playing a Richard II Shakespeare play [4] and a synchronised companion screen application on a mobile device that engages users through different levels of interaction.

Companion Screen Application

The companion application was an interactive and synchronised transcript of a Shakespeare play. The concept was inspired by theatre *surtitles* which aim at improving the comprehensibility of opera libretti and plays presented in foreign or ancient languages. The user interface design adopts the layout of a textbook. This presentation layout allowed the viewer to read text lines presented by the scene characters, as well as reading ahead. The shape of a tablet fits the aspect ratio of a book and we feel it inherently supports the textbook metaphor. The companion application was designed to operate in three modes to enable the three interaction levels which we planned to study: *passive*, *exploration* and *call-to-action*. Figure 1 shows screen shots of the graphical user interface (GUI) in all three modes.

Levels of Interaction

In *passive* mode, the GUI showed the synchronised text but did not respond to user input. Participants were only able to consume the companion content and had no control over the presentation. In this mode, the GUI comprised of the following elements: a menu bar and a container displaying the synchronised transcript. The menu bar showed icons containing pictures of the characters who had active roles (lines) in the current scene. The icon corresponding to the current speaker was highlighted. Highlighting was achieved by enlarging the icon and changing its border colour. The line that was being spoken on the TV was highlighted in the textbook within the paragraph that contained the line. The current paragraph was highlighted by a vertical black bar on the left hand side of the paragraph. The current line was emphasised by changing its background colour.

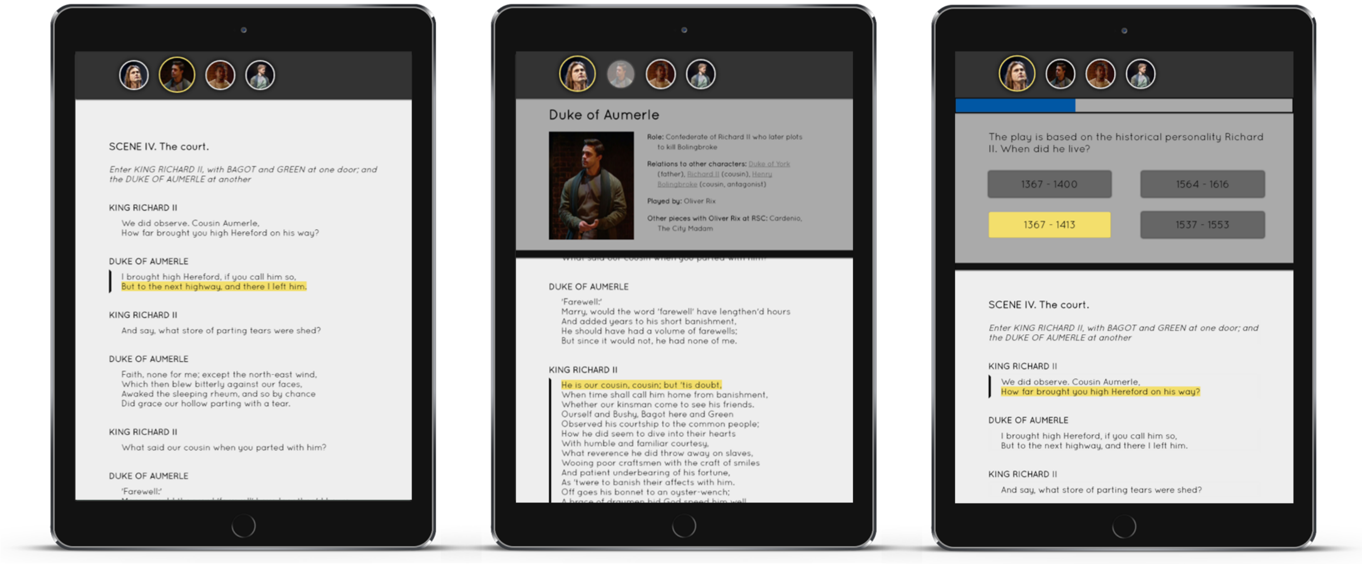
The *exploration* mode provided a casual form of interaction with the companion content. The character icons in the menu bar were linked and activated. When an icon was clicked, a menu appeared from the top of the screen to offer information on the relevant character. The character information comprised of the characters name, their picture, their role in the play, relationships to other characters in the play, the actors name and other productions of the Royal Shakespeare Company the actor starred in. The information menu was presented as an overlay (middle image in Figure 1). To ensure that the highlighted line was still visible, the text container got repositioned with the highlighted line in the vertical centre of the remaining visible part of the text container.

In the *call-to-action* mode, the participant was presented with a play-along experience in the form of quiz questions on the tablet. The quiz questions were not directly relevant to the exact scene shown on the TV. Instead, they were related to the general subject of the play (Richard II [4]) and Shakespeare in general. Users needed to select one of four responses to the question within 20 seconds. The time left to select an answer was indicated by a progress bar on top of the quiz menu. Like the character information menu, the quiz menu appeared from the top of the screen. An accompanying flash notification (the call-to-action) appeared on the bottom right corner of the TV to inform the user about the new question on the companion screen.

Synchronisation

Synchronisation between the companion and the TV content was achieved by means of the DVB-CSS suite of protocols [7, 8], which are also part of the HbbTV 2.0 specification [6]. First implementations of the protocol have been shown to allow frame accurate synchronisation between devices [28]. Unfortunately, at the time of this line of experimentation, no TV sets implementing the DVB-CSS protocols were publicly available. Thus a TV emulator based on an open source Python implementation [12] of the DVB-CSS protocols was used. The front-end of the emulator consisted of a JavaScript-based Web application which instantiated the Shakespeare video in a full screen video player.

The companion application consisted, similar to the one presented by Vinayagamoorthy et al. [28], of a DVB-CSS client on an iOS device. It connected the companion screen to the protocol endpoints of the TV emulator to receive updates of the TV's video presentation time. The user interface of the companion application was implemented as a web application running in an embedded browser (web view). The compan-



SCENE IV. The court.

Enter KING RICHARD II, with BAGOT and GREEN at one door; and the DUKE OF AUMERLE at another

KING RICHARD II

We did observe. Cousin Aumerle,
How far brought you high Hereford on his way?

DUKE OF AUMERLE

I brought high Hereford, if you call him so,
But to the next highway, and there I left him.

KING RICHARD II

And say, what store of parting tears were shed?

DUKE OF AUMERLE

Faith, none for me; except the north-east wind,
Which then blew bitterly against our faces,
Awaked the sleeping rheum, and so by chance
Did grace our hollow parting with a tear.

KING RICHARD II

What said our cousin when you parted with him?

DUKE OF AUMERLE

'Farewell:'

Duke of Aumerle

Role: Confederate of Richard II who later plots to kill Bolingbroke

Relations to other characters: Duke of York (father), Richard II (cousin), Henry Bolingbroke (cousin, antagonist)

Played by: Oliver Rix

Other pieces with Oliver Rix at RSC: Cardenio, The City Madam

DUKE OF AUMERLE

'Farewell:'
Marry, would the word 'farewell' have lengthen'd hours
And added years to his short banishment,
He should have had a volume of farewells;
But since it would not, he had none of me.

KING RICHARD II

He is our cousin, cousin; but 'tis doubt,
When time shall call him home from banishment,
Whether our kinsman come to see his friends.
Ourself and Bushy, Bagot here and Green
Observed his courtship to the common people;
How he did seem to dive into their hearts
With humble and familiar courtesy,
What reverence he did throw away on slaves,
Wooing poor craftsmen with the craft of smiles
And patient underbearing of his fortune,
As 'twere to banish their affects with him.
Off goes his bonnet to an oyster-wench;

The play is based on the historical personality Richard II. When did he live?

SCENE IV. The court.

Enter KING RICHARD II, with BAGOT and GREEN at one door; and the DUKE OF AUMERLE at another

KING RICHARD II

We did observe. Cousin Aumerle,
How far brought you high Hereford on his way?

DUKE OF AUMERLE

I brought high Hereford, if you call him so,
But to the next highway, and there I left him.

KING RICHARD II

And say, what store of parting tears were shed?

Figure 1. Screen-shots of the companion application in modes of interaction: passive (left), exploration (middle) and call-to-action (right)

ion's estimates of the TV's presentation time are reported from the DVB-CSS client to the web application.

The time which elapses from the DVB-CSS client sending a timing update and the web view rendering a corresponding presentation to companion screen is unknown. To determine this duration, the DVB-Sync-Timing Framework [13] was used. The result was then used as a timing-calibration offset in the web application.

The textbook application on the companion device and the video on the TV are defined to be in sync, if the highlight on a text line in the textbook application is set at the same point in time as the character on the TV starts speaking the corresponding line. Content timings, the points on the video presentation timeline at which a certain text line on the companion shall be highlighted was authored to allow the best possible control over the synchronisation cues.

To do this, we extracted the audio stream from the TV content and imported it into a digital audio workstation (DAW), which supports visualisation of the audio waveform. To measure the timing of a text line, the play position marker was placed at the beginning of the waveform corresponding to this line. The position of the marker is read from DAW's play position display. Timings are recorded in JSON format for later interpretation by the companion application. The authoring process is illustrated in Figure 2.

EXPERIMENT METHODOLOGY

In this section, we describe the experimental set-up used to evaluate hypotheses defined above.

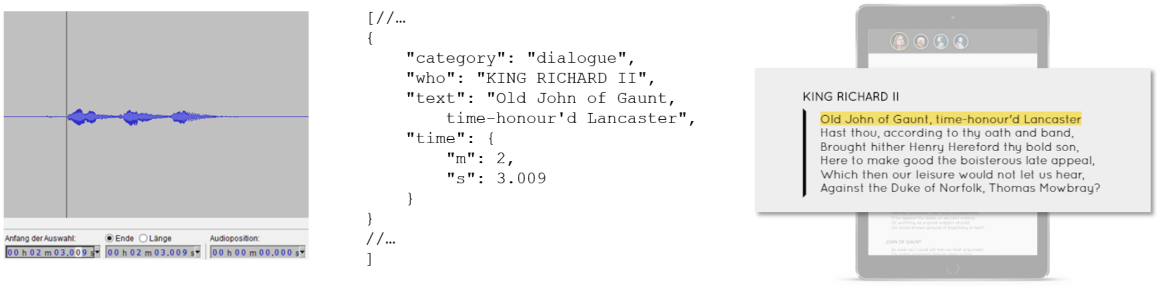
Experiment Variables

During the experiment values of two independent variables are varied. These were *a)* the delay between the content on the TV and the companion device and *b)* the type of interaction. Table 1 summarises the choice of variables.

The choice of delay levels is based on results of a pilot experiment with engineers working in the broadcast sector who were skilled at noticing delays in synchrony. In the pilot, we tested delays in the order of magnitude of perceived lip sync (±50 ms and ±100 ms) [26], of the average speech rate (±200 ms, ±500 ms) [29] as well as values in the order of magnitude of accepted delays for TV subtitles ±1000 ms and 2000 ms [5, 17]. Positive delays indicate that the tablet content is behind the TV content, negative values that tablet content is ahead. As only a few of the participants noticed the delays in lip-synchronism range, the delay values ±50 ms and ±100 ms were excluded from the set of factor levels to study in the main experiment. In order to reduce the size of the experiment, the largest negative and the largest positive delays were removed. The subset of participants who noticed the positive delay in the pilots were the same for 1000 ms and 2000 ms. In addition, previous research on media synchronisation showed a symmetrical effect for large negative and positive delays [26]. The final set of time delay levels used in the main experiment was composed of the reference control condition 0 ms and the five delay values ±250 ms, ±500 ms and 1000 ms.

Three levels of interaction were defined: *passive*, *exploration* and *call-to-action*. The interaction level 'passive' was defined to be the lowest interaction level, as participants have no way to interact other than to follow the rolling text. Level 'call-to-action' was defined to be the highest level of interaction, as the application engaged the participant to interact, while at level 'exploration', participants were free to interact with the companion content.

Two objective measures were applied to record participants responses. To gain data which might give us an insight into the participants gaze behaviour (attention split), we annotated



Anfang der Auswahl: 00 h 02 m 03.009 s

● Ende ○ Länge 00 h 02 m 03.009 s

Audioposition: 00 h 00 m 00.000 s

```
[//…
{
    "category": "dialogue",
    "who": "KING RICHARD II",
    "text": "Old John of Gaunt,
        time-honour'd Lancaster",
    "time": {
        "m": 2,
        "s": 3.009
    }
}
//…
]
```

KING RICHARD II

Old John of Gaunt, time-honour'd Lancaster
Hast thou, according to thy oath and band,
Brought hither Henry Hereford thy bold son,
Here to make good the boisterous late appeal,
Which then our leisure would not let us hear,
Against the Duke of Norfolk, Thomas Mowbray?

Figure 2. Authoring of text timings: amplitude of the audio recording in the DAW, corresponding chunk of the JSON representation of the text timings and presentation of the highlighted text line in the companion application.

participants gaze while watching the experiences through the use of video recordings taken during the experiment. Obviously, gaze recordings do not mirror what participants actually think but they provided us with an indication of people's visual attention. In prior work, Neat et al. [23] used gaze recordings as an objective measure in an experiment as a means for mediating viewers attention between a companion screen and a TV. In their analysis, they correlated gaze changes with different triggers to assess the effectiveness of these triggers to drive the users attention between screens. We chose to derive two values from our gaze recordings: the *dwell* on tablet p_{tablet} and the gaze change frequency f_{focus}. p_{tablet} is the share of time the user looks at the companion whilst being exposed to an experiment condition. f_{focus} tells us how often users' gaze switches between the TV and companion screen.

To gain information on how participants interact with the companion application, we logged all interaction (button presses, swipe gestures) into a database. From the interaction logs we compute the interaction frequency $f_{interact}$, which tells us about how participants interact while being exposed to an experiment condition.

As a subjective measure, participants ratings on questions from a questionnaire were recorded. A 7-point Likert type scale, as recommended by [9], was chosen to capture ratings. Questions cover perception of delays (Q1 to Q3) and attention split (Q4 and Q5) between companion and TV:

Q1 I felt the TV and the Tablet were well timed to each other.

Q2 I felt like I was waiting for the TV to "catch up" to the Tablet.

Q3 I felt like I was waiting for the Tablet to "catch up" to the TV.

Q4 I felt I was missing part of the programme on the TV because of the Tablet.

Q5 I felt I was missing content on the Tablet because of the TV.

Previous QoE-centric studies on media sync asked participants if they noticed a delay and whether this delay was annoying, for example [20, 21, 26]. This approach might prime users to watch out for delays and suggest that delays affect QoE. The design of Q1 to Q3 aimed at avoiding these effects. Q4 and Q5 where derived from a questionnaire by Neate et al. [22], which they used in a study on attention split between TVs and companion screens.

Setup

The study was planned as a repeated-measures within-participant design. All subjects were exposed to all conditions including reference conditions. As there were six factor levels for the delay and three levels of interaction, there were 18 conditions each participant experienced.

To eliminate the influence of order-effects, conditions were counterbalanced across subjects. This was achieved by dividing participants into three groups of six participants. Each participant was presented three blocks of six clips. Each block was assigned to one interaction level. Within one block, each participant saw all delay levels. Participants within a group were presented the same sequence of interaction levels. Within one block of a group, delay levels were counterbalanced by means of 6×6 latin squares [3]. To prevent an effect of the order in which participants were exposed to the same interaction levels, the interaction level is varied across the three groups of participants in a 3×3 latin square.

Combinations of delay levels and experiences were applied to 18 different clips of a recording of the Shakespeare play Richard II performed at the Royal Shakespeare Company. Clips were chosen with equal amount of dialogue and scene complexity to eliminate the influence of the clip order or type on the measurement. However, the order of clips was randomised to avoid a specific semantic relation between subsequent clips.

Before the experiment started participants were shown a demo clip (at zero delay and at interaction level passive) so they got acclimatised to the new experience. Before being exposed to each block of six clips, participants were briefed on the type of interaction. After each clip, participants were asked to answer questions Q1 to Q5. After 12 clips participants were asked to take a 15 minutes break to prevent viewing fatigue. Clips had an average length of 90 seconds. In general, participants spent about one minute to fill the questionnaire in between conditions. Including welcome, briefings, collecting questionnaire responses, break and debrief one experiment session took about 120 minutes.

Name	Domain	Unit	Hypothesis
Independent variables			
Delay (d)	$\{-500, -250, 0, 250, 500, 1000\}$	ms	all
Level of interaction (e)	$\{$P,I,Q$\}$	-	all
Dependent Variables			
Questionnaire rating (r)	$\mathbb{N} \wedge 1 \leq r \leq 7$	-	H1, H5
Dwell on tablet (p_{tablet})	$\mathbb{R} \wedge -1 \leq p_{tablet} \leq 1$	min/min	H2
Gaze change frequency (f_{focus})	$\mathbb{R} \wedge 0 \leq f_{focus} < \infty$	$1/min$	H3
Interaction frequency ($f_{interact}$)	$\mathbb{R} \wedge 0 \leq f_{interact} < \infty$	$1/min$	H4

Table 1. Independent and dependent variables and related hypotheses (interaction levels: passive *P*, explorative *I*, call-to-action *Q*).

Experiments were conducted in a user experience lab arranged to look as close to a natural TV viewing environment. It contained a TV on a sideboard along with peripheral devices like a set-top box and gaming consoles. It was surrounded by sofas and a coffee table. The room was equipped with cameras and microphones to observe the participants during the experiment. Recordings were used after the experiment for gaze observations as shown in Figure 3.

Participants

The 18 participants recruited for the experiment were chosen to represent the population of potential users of the companion screen application. This included requirements with regard to affinity for Shakespearean plays or theatre in general, TV viewing behaviour, tablet or smart phone usage. Recruitment was done by a specialised agency. Participants received an incentive of about the equivalence of € 70.00. We also collected demographics data and media usage behaviour from participants.

Participants were between 18 and 55 years of age (M=38, SD=13) and gender balanced. The youngest participant was 18 years of age and the oldest was 55. Participants watched on average 2.26 hours of TV per day (SD=1.95 hours, Max=9 hours, Min=0.5 hours). Participants were asked to indicate their perceived level of computer literacy. 14 participants (77.78%) self-reported a rating of $\geq$5 on a 7-point Likert type scale and 3 participants (5.56%) estimated their computer literacy $\leq$3. Asked on whether they liked watching Shakespeare plays, 10 participants (55.56%) gave a rating $\geq$5 and 2 participants (11.11%) a rating of $\leq$3. 17 participants (94.44%) gave a rating of $\geq$5 when asked if they often used a touch-screen device such as tablet or smart phone. No participant gave a rating of $\leq$3. All participants stated that they used a secondary device (smart phone, tablet, laptop) to search for information related to the TV programme.

In summary, all participants appeared to be familiar with the technology used in the experiment. The majority of participants assessed themselves as frequent users of touch-screen devices and feel familiar using a computer. All participants had experience of using a secondary device whilst watching television and most of the participants stated an interest in Shakespeare plays.

RESULTS

Delay Perception Threshold (H1)

Participants responses to questions Q1, Q2 and Q3 were analysed to test hypothesis H1. Single-factor analysis was conducted across the responses captured for each of the delay values at the controlled interaction level. This allowed for finding the delay-perception tolerance at each interaction level. Shapiro-Wilk tests showed that samples could not be assumed to come from normally distributed populations. Therefore, the non-parametric Friedman test was applied to test the influence of the delay on the participants ratings. Table 2 depicts the corresponding test results. A significant effect of the delay level on the participants rating was not found for any of the questions on synchronisation accuracy. This result is also outlined by Figure 4, which shows similar medians and quartile ranges of ratings on Q1 across the different delay values at all levels of interaction. The delay-tolerance level was not determined within this study. Thus there is not enough information for a final judgement on H1.

e	Dependent variable	χ_r^2	n	p	H_0
P	r_3	4.28	18	0.51	accepted
	r_4	5.23		0.39	accepted
	r_5	5.9		0.32	accepted
I	r_3	1.9	17	0.83	accepted
	r_4	3.96		0.56	accepted
	r_5	3.08		0.69	accepted
Q	r_3	9.81	17	0.09	accepted
	r_4	6.8		0.24	accepted
	r_5	7.12		0.22	accepted

Table 2. Tests on hypothesis H1: Results of the Friedman test show no significant influence of the delay on the subjects' ratings r on questions on synchronisation accuracy Q1, Q2 and Q3 for different levels of interaction e (P passive, I explorative, Q call-to-action) ($df = 5$).

Dwell (H2) and Gaze-Change Frequency (H3)

For an assessment of H2, the influence of the delay on the participants' mean dwell on the tablet was analysed. Shapiro-Wilk tests showed that samples could be assumed to come from normally distributed populations. Hence, a parametric test, the repeated-measures single-factor analysis of variance (rANOVA), was applied to analyse the effect of the delay. Results are presented in Table 3. A significant effect caused by the delay was not found across the interaction levels. Hence, results do not confirm H2. p-values were calculated under the assumption of sphericity. The results of the Mauchly test for sphericity are also provided in Table 3. For each interaction level, mean ratings were at similar levels across delay levels ($min(\mu_{p_{tablet}}(P)) = 0.39$ vs. $max(\mu_{p_{tablet}}(P)) = 0.48$, $min(\mu_{p_{tablet}}(I)) = 0.52$ vs. $max(\mu_{p_{tablet}}(I)) = 0.57$, $min(\mu_{p_{tablet}}(Q)) = 0.56$ vs. $max(\mu_{p_{tablet}}(Q)) = 0.59$), which is inline with the results of the hypothesis test.

To test H3, the influence of the delay on the participants' relative gaze-change frequency was analysed. A Shapiro-Wilk test showed that samples could be assumed to come from a

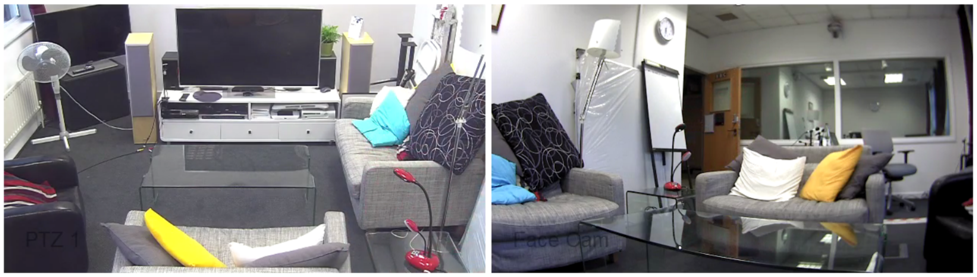


Figure 3. Lab: left picture is taken from behind the participants seat. Right picture is taken from a camera mounted next to the television and shows the participants seat. The latter picture is used for the gaze observation.

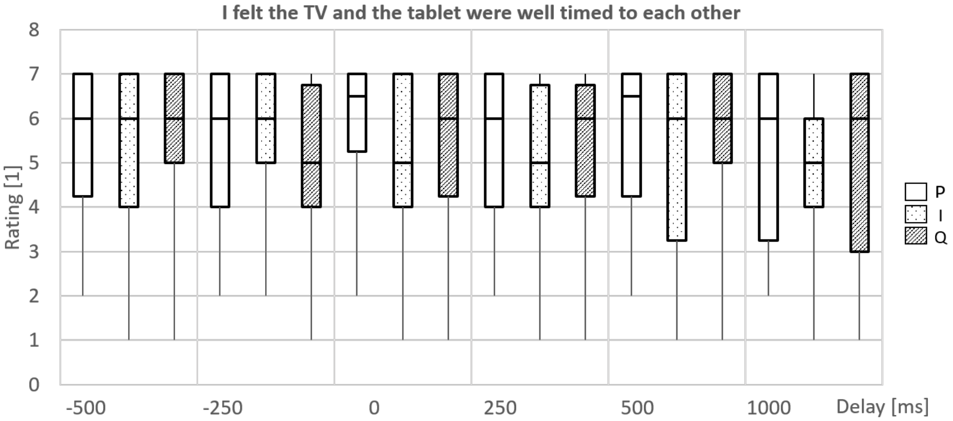


Figure 4. Box plots off participants' ratings on Q1 for different levels of interaction.

normally distributed population. Thus, rANOVA was applied. p-values were calculated under the assumption of sphericity. A significant impact of the delay was found for interaction level *exploration*. As H3 hypothesises claims a higher f_{focus} at higher delay levels, a one-tailed t test was used for pairwise comparison of samples applying Bonferroni correction of p-values. The results are presented in Table 4. Row "Expected" in Table 4 marks those pairs of samples, where H_0 (with alternative hypothesis $H_1 : \mu_{d_1} < \mu_{d_2}$) is expected to be rejected in order to support H3. Results show that f_{focus} was significantly larger at a delay level of -500 ms ($\mu(-500) = 19.82\,\text{min}^{-1}$, $S(-500) = 5.59\,\text{min}^{-1}$) as opposed to -250 ms ($\mu(-250) = 16.08\,\text{min}^{-1}$, $S(-250) = 5.93\,\text{min}^{-1}$).

However, no such effect was found between sample pairs at the reference condition 0 ms and the largest negative or positive delay values -500 ms and 1000 ms respectively, or any other of the sample pairs that were expected to show an effect. Also the mean value of f_{focus} at 250 ms ($\mu(-250) = 21.40\,\text{min}^{-1}$, $S(-250) = 9.82\,\text{min}^{-1}$) was found to be significantly larger than at -250 ms, though both delay levels have the same absolute values. An effect of the delay on f_{focus} was observed. Grounded on Cohen's f the effect size is small to medium ($f = 0.167$). However, the direction of the effect of the delay on f_{focus} claimed by H3 was not confirmed by the experiment results. An additional post-hoc test was conducted to find out, if there was complementary direction of the effect ($H_1 : \mu_{d_1} < \mu_{d_2}$). The results are provided in Table 5. An effect was only found between the reference level ($\mu(0) = 20.02\,\text{min}^{-1}$, $S(0) = 7.32\,\text{min}^{-1}$) and -250 ms. A complementary direction of the effect of the delay on f_{focus} is, therefore, not confirmed.

Influence of the delay on interaction frequency (H4)

To evaluate hypothesis H4, the influence of the delay on the interaction frequency was tested at interaction levels *exploration* and *call-to-action*. Shapiro-Wilk tests showed that the majority of the samples could not be assumed to come from normally distributed populations. Thus, the Friedman test was applied. Results are presented in Table 6. No significant effect of the delay on the interaction frequency was found. For interaction level Q, median values vary within a narrow range of $min(\theta_{f_{interact}}) = 0,70\,\text{min}^{-1}$ and $max(\theta_{f_{interact}}) = 0,81\,\text{min}^{-1}$. This observation is in line with the result of the Friedman test. For interaction level I medians vary within a larger range of $min(\theta_{f_{interact}}) = 0,80\,\text{min}^{-1}$ and $max(\theta_{f_{interact}}) = 2.20\,\text{min}^{-1}$. The larger range of medians is also reflected in the lower p value derived from the Friedman test. However, this deviation can not be regarded as significant. The results of the experiment do not deliver evidence to support H4.

Feeling of being distracted (H5)

Validity of hypothesis H5 is assessed by analysis of the influence of the delay on participants ratings on questions Q4 and Q5 at different levels of interaction. A Shapiro-Wilk test showed that samples could not be assumed to come from normally distributed populations. Hence, the Friedman test was applied. Results of the Friedman test are shown in Table 7.

e	Dependent variable	Notes on sphericity	F	p	H_0	η^2
P	p_{tablet}	SpA (MW=0.23, $p = 0.12$)	0.65	0.66	accepted	0.045
	f_{focus}	SpA (MW=0.20, $p = 0.07$)	0.73	0.61	accepted	0.043
I	p_{tablet}	SpA (MW=0.37, $p = 0.46$)	0.44	0.81	accepted	0.027
	f_{focus}	SpA (MW=0.43, $p = 0.63$)	0.25	0.94	accepted	0.016
Q	p_{tablet}	SpA (MW=0.19, $p = 0.06$)	0.39	0.86	accepted	0.024
	f_{focus}	SpA (MW=0.33, $p = 0.35$)	3.13	0.01	rejected	0.164

Table 3. Tests on hypotheses H2 and H3: Results of the rANOVA on influence of the delay on p_{tablet} and f_{focus} for interaction levels ($df_{treatment} = 5$, $df_{residuals} = 80$, "SpA": sphericity assumed), "MW": Mauchly-W.

d_1 [ms]	0					-250				-500			250		500
d_2 [ms]	-500	-250	250	500	1000	-500	250	500	1000	250	500	1000	500	1000	1000
Expected	*	*	*	*	*	*		*	*			*	*	*	*
p	1	1	1	1	1	0.04	0.04	0.45	1	1	1	1	1	1	1
H_0	A	A	A	A	A	R	R	A	A	A	A	A	A	A	A

Table 4. Post-hoc analysis for hypothesis H3: pairwise one-tailed ($H_1 : \mu_{d_1} < \mu_{d_2}$) T tests with Bonferroni-corrected p-values on samples of f_{focus} measured at different delay levels at interaction level call-to-action. Columns, where rejection of H_0 is expected in order support H3 are marked with "*". Columns where H_0 is accepted or rejected are marked with "A" and "R" respectively.

The Friedman test shows a significant effect of the delay on the ratings on question Q4.

A post-hoc analysis, to determine which sample pairs caused the significant effect, was conducted by means of a pairwise one-tailed ($H_1 : \pi^+_{d_1} < 0.5$) binomial-sign test for dependent samples. Results are presented in Table 8. They show that a significantly higher amount of participants gave a higher rating on question Q4 at delay level 1000 ms as opposed to the reference condition 0 ms ($\sum D^+ = 10, \sum D^- = 2$) and the delay levels -500 ms ($\sum D^+ = 10$, $\sum D^- = 2$), 250 ms ($\sum D^+ = 7$, $\sum D^- = 1$) and 500 ms ($\sum D^+ = 9$, $\sum D^- = 2$). Moreover, a significant amount of participants gave a higher rating on Q4 at 250 ms as opposed to -500 ms ($\sum D^+ = 7, \sum D^- = 2$). Figure 5 illustrates these findings. Red circles mark median values of those samples where an effect was found. Arrows connect the pairs of samples for which an effect was found. The arrow points into the direction of the sample with significant amount of lower ratings. An additional one-tailed test in opposite direction ($H_1 : \pi^+_{d_1} > 0.5$) did not find significant differences between samples.

DISCUSSION

We did not see a significant difference, in participants ratings of questions Q1 to Q3, across the different time delays (-500 ms to 1000 ms) for any of the chosen interaction levels. This implies that, for the type of experience discussed in this paper, our data set does not uncover a time delay at which participants are able to perceive inaccurate synchronisation for any of the three different interaction types designed for the companion application used. Therefore, our data does not deliver the information basis for a final judgement on hypothesis H1.

However, the results implies that, for the type of companion experience described in this paper, synchronisation delays between -500 ms and 1000 ms are unlikely to be noticed. This in turn suggests that prevalent technologies like Audio Watermarking or Audio Fingerprinting, which are reported to achieve synchronisation accuracy within a fraction of a second [11, 27], and DVB-CSS [7, 8, 28] are viable technological solutions for implementing companion experiences similar to the ones tested in this paper. This also requires that errors introduced during production of content timings and errors introduced by the synchronisation system, do not accumulate to a value that exceeds the investigated delay range. Manual crafting of timings for the Shakespearean play transcript, used in the study, was comparatively time consuming. A production process that is at least partially automated will be a more practicable and efficient process. Future work could investigate efficient ways to generate the timing information, in particular the applicability of professional subtitling tools or natural language processing tools. For the specific application used in our work, a major challenge in this regard is certainly the processing of the Shakespearean language.

Hypotheses on the impact of the delay on dwell time H2, gaze change frequency H3 and interaction frequency H4 were not confirmed by the experiment. However, we observed a significant effect, across the time delay, when users were able to actively browse content on the companion screen (interaction level - exploration). This shows that delays can effect aspects of the user experience before users start noticing them.

This finding has implications on the method used to evaluate a companion application before going "live". Authors follow a certain intention, when orchestrating content across screens, for example to mediate users attention between companion and TV screens, as shown by Neate et al. [23]. Tests that only look at perceptibility of delays, may not suffice to evaluate whether a specific synchronisation system is able to support the authors' intent.

CONCLUSION

We investigated the perception of delays, between content presentation on companion screen devices and TVs, as well as the impact of delays on some aspects of the user experience. Our results indicate that, for companion screen applications which aim to present users with a synchronised rolling transcript as an accompaniment to a programme with relatively difficult subject matters (such as the Shakespearean play used in this study), synchronisation technologies are able to keep delays below the threshold at which users may perceive mistimed content across devices. With one exception, we saw no effect of the chosen delay levels in the range of $[500ms, 1000ms]$ on

d_1 [ms]	0					-250				-500			250		500
d_2 [ms]	-500	-250	250	500	1000	-500	250	500	1000	250	500	1000	500	1000	1000
p	1	0.01	1	1	0.43	1	1	1	1	1	1	0.82	1	0.34	1
H_0	A	R	A	A	A	A	A	A	A	A	A	A	A	A	A

Table 5. Post-hoc analysis for hypothesis H3: Test on complementary direction of the effect of the delay on f_{focus} as claimed by H3: pairwise one-tailed ($H_1 : \mu_{d_1} > \mu_{d_2}$) T tests with Bonferroni-corrected p-values on samples of f_{focus} measured at different delay levels at interaction level call-to-action. Columns where H_0 is accepted or rejected are marked with "A" and "R" respectively.

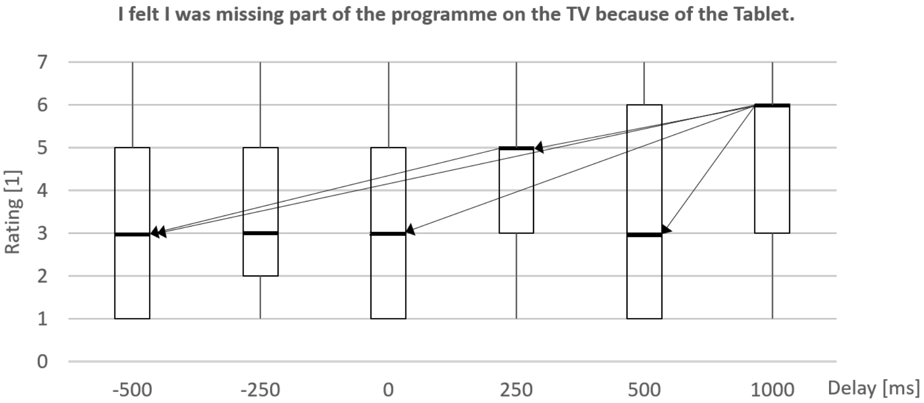


Figure 5. Box plots of ratings on Q4 for different levels of delay at interaction level exploration. Arrows connect pairs of samples for which an effect was found in the pairwise sign test as shown in Table 8. The arrows point into the direction of the samples with significant amount of lower ratings.

e	Dependent variable	χ_r^2	n	p	H_0
I	$f_{interact}$	8.86	16	0.12	accepted
Q	$f_{interact}$	5.29	14	0.39	accepted

Table 6. Test on hypothesis H4: results of the Friedman test on the influence of the delay ($df = 5$) on the interaction frequency for different levels of interaction e.

e	Dependent variable	χ_r^2	p	H_0
P	r_5	4.90	0.43	accepted
	r_6	8.16	0.15	accepted
I	r_5	11.56	0.04	rejected
	r_6	6.42	0.27	accepted
Q	r_5	9.64	0.09	accepted
	r_6	5.79	0.33	accepted

Table 7. Test on hypothesis H5: results of the Friedman test ($df = 5$, $n = 17$) on the influence of the delay on subjects' ratings r on questions Q4 and Q5 for different levels of interaction e.

participants responses. During the conditions in which participants were able to actively browse additional content on the companion-screen, users' felt more distracted by the tablet at a delay of 1000 ms as opposed to lower delay values. The participants lack of noticing any delays within the given range may be an indication that evaluation methods focussing on users perception of delays may not suffice to assess the overall requirements on timing accuracy to support specific goals in user experience design.

ACKNOWLEDGMENTS

We would like to thank the participants who took part in our experiment, our colleagues who participated in pilot trials and the reviewers for helping us improve this paper. The research leading to these results has received funding from the European Union's H2020-ICT-2015 programme under grant agreement n° 687655 (2-IMMERSE).

REFERENCES

1. Mike Armstrong. 2013. *The Development of a Methodology to Evaluate the Perceived Quality of Live TV Subtitles*. White Paper WHP 259. British Broadcasting Cooperation.

2. Ingar M. Arntzen, Njål T. Borch, and Christopher P. Needham. 2013. The media state vector: a unifying concept for multi-device media navigation. In *MoVid '13 Proceedings of the 5th Workshop on Mobile Video*. 61–66.

3. R. A. Bailey. 2008. *Design of Comparative Experiments*. Cambridge University Press, Chapter Row-column designs, 105–116.

4. British Broadcasting Cooperation. 2016. Richard II. `http://www.bbc.co.uk/programmes/p03rr1v1`. (2016). Online; accessed: 2016-08-27.

5. D. Burnham, J. Robert-Ribes, , and R. Ellison. 1998. Why captions have to be on time. In *Audio-visual speach processing*. 153–156.

6. European Telecommunications Standards Institute (ETSI) 2015a. *ETSI TS 102 796 V1.3.1 / HbbTV 2.0 – Hybrid Broadcast Broadband TV* . European Telecommunications Standards Institute (ETSI).

7. European Telecommunications Standards Institute (ETSI) 2015b. *ETSI TS 103 286-1 – Companion Screens and Streams; Part 1: Concepts, roles and overall architecture* (DVB BlueBook A167-1 ed.). European Telecommunications Standards Institute (ETSI).

8. European Telecommunications Standards Institute (ETSI) 2015c. *ETSI TS 103 286-2 – Companion Screens and Streams; Part 2: Content Identification and Media Synchronization* (DVB BlueBook A167-2 ed.). European Telecommunications Standards Institute (ETSI).

d_1 [ms]	0					-250				-500			250		500
d_2 [ms]	-500	-250	250	500	1000	-500	250	500	1000	250	500	1000	500	1000	1000
Expected	*	*	*	*	*	*		*	*			*	*	*	*
p	0.37	0.37	0.1	1	0.02	1	0.37	1	0.1	0.02	1	0.02	1	0.002	0.02
H_0	A	A	A	A	R	A	A	A	A	R	A	R	A	R	R

Table 8. Post-hoc analysis for hypothesis H5: results of pairwise one-tailed binomial-sign test ($H_1 : \pi^{+}_{d_1} < 0.5$) on responses to Q4 at interaction level explorative. P values are Bonferroni corrected. Columns, where rejection of H_0 is expected in order support H5 are marked with "*". Columns where H_0 is accepted or rejected are marked with "A" and "R" respectively.

9. Kraig Finstad. 2010. Response Interpolation and Scale Sensitivity: Evidence Against 5-Point Scales. *Journal of Usability Studies* 5, 3 (2010), 104–110.

10. David Geerts, Ishan Vaishnavi, Rufael Mekuria, Oskar van Deventer, and Pablo Cesar. 2011. Are we in Sync? Synchronization Requirements for Watching Online Video Together. In *CHI '11 Conference on Human Factors in Computing Systems*. 311–314.

11. Leandro Gomes, Pedro Cano, Emilia Gomez, Madeleine Bonnet, and Eloi Batlle. 2003. Audio Watermarking and Fingerprinting: For Which Applications? *Journal of New Music Research* 32, 1 (2003), 65–82.

12. Matt Hammond. 2016. Python DVB Companion Screen Synchronisation protocol library. `https://github.com/BBC/pydvbcss`. (2016). Online; accessed: 2016-08-08.

13. Matt Hammond and Jerry Kramskoy. 2016. DVB companion synchronisation timing accuracy measurement. `https://github.com/bbc/dvbcss-synctiming`. (2016). Online; accessed: 2016-08-24.

14. Nielsen Holdings. 2011. In the U.S., tablets are TV buddies while eReaders make great bedfellows. `http://www.nielsen.com/us/en/insights/news/2011/in-the-u-s-tablets-are-tv-buddies-while-ereaders-make-great-bedfellows.html`. (2011). Online; accessed: 2016-09-02.

15. International Organization for Standardization (ISO) 2010. *ISO 9241-210:2010: Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems*. International Organization for Standardization (ISO).

16. Internet Engineering Task Force (IETF) 2014. *RFC 7272 – Inter-Destination Media Synchronization (IDMS) Using the RTP Control Protocol (RTCP)* (Request for Comments 7272 ed.). Internet Engineering Task Force (IETF).

17. Ichiro Maruyama, Yoshiharu Abe, Eiji Sawamura, Tetsuo Mitsuhashi, Terumasa Ehara, and Katsuhiko Shirai. 1999. Cognitive experiments on timing lag for superimposing closed captions. In *Sixth European Conference on Speech Communication and Technology*. 575–578.

18. Sebastian Moeller and Alexander Raake. 2014. *Quality of Experience – Advanced Concepts, Applications and Methods*. Springer, Chapter Quality and Quality of Experience, 11–33.

19. Mario Montagud, Fernando Boronat, Hans Stokking, and Ray van Brandenburg. 2012. Inter-destination multimedia synchronization: schemes, use cases and standardization. *Multimedia Systems* 18, 6 (2012), 459–482.

20. Mu Mu, Steven Simpson, Hans Stokking, and Nicholas Race. 2016. QoE-aware Inter-stream Synchronization in Open N-Screens Cloud. In *13th Annual IEEE Consumer Communications & Networking Conference (IEEE CCNC)*. 907–915.

21. Niall Murray, Yuansong Qiao, Brian Lee, A. K. Karunakar, and Gabriel-Miro Muntean. 2013. Subjective evaluation of olfactory and visual media synchronization. In *4th ACM Multimedia Systems Conference (MMSys '13)*. 162–171.

22. Timothy Neate, Michael Evans, and Matt Jones. 2016. Designing Visual Complexity for Dual-screen Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 475–486.

23. Timothy Neate, Matt Jones, and Michael Evans. 2015. Mediating Attention for Second Screen Companion Content. In *CHI '15 Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3103–3106.

24. Ofcom. 2016. Communications Market Report 2016. `https://www.ofcom.org.uk/__data/assets/pdf_file/0024/26826/cmr_uk_2016.pdf`. (2016). Online; accessed: 2017-03-28.

25. Red Bee Media Ltd. 2012. Broadcast industry not capitalising on rise of the second screen. `http://www.redbeemedia.com/sites/all/files/downloads/secondscreenresearch.pdf`. (2012). Online; accessed: 2013-07-15.

26. Ralf Steinmetz. 1996. Human perception of jitter and media synchronization. *IEEE Journal on Selected Areas in Communications* 14, 1 (1 1996), 61–72.

27. Vinoba Vinayagamoorthy, Matt Hammond, Penelope Allen, and Michael Evans. 2012. Researching the User Experience for Connected TV – A Case Study. In *CHI EA Extended Abstracts on Human Factors in Computing Systems*. 589–604.

28. Vinoba Vinayagamoorthy, Rajiv Ramdhany, and Matt Hammond. 2016. Enabling Frame-Accurate Synchronised Companion Screen Experiences. In *TVX '16 Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*. 83–92.

29. J. Yuan, M. Liberman, and C. Cieri. 2006. Towards an integrated understanding of speaking rate in conversation. In *Interspeech 2006*. 541–544.