Assessment of Technical Quality of Online Video Using Visualization in Place of Experience

By

Diba Kaya

A thesis submitted in conformity with the requirements for the degree of Masters of Applied Science in Industrial Engineering

Industrial and Mechanical Engineering Graduate Department

© Copyright by Diba Kaya, 2016
Assessment of Technical Quality of Online Video Using Visualization in Place of Experience

Diba Kaya
Masters of Applied Science
Industrial and Mechanical Engineering
University of Toronto
2016

Abstract
This thesis introduces a new method to measure subjective ratings of technical quality (TQ) in disrupted video. Current methods are subject to extraneous factors such as delays introduced by the streaming software and video content. The proposed method is designed to supplement current methods by avoiding the effects of content on TQ ratings and reducing the time and effort necessary to run experiments. The new method replaces actual videos with images representative of videos that contain disruptions. I ran an experiment with 37 participants to explore the viability of assessing TQ using visualizations instead of actual videos. The experiment compared ratings made after watching videos that varied in TQ and ratings made after viewing TQ visualizations only. The results of the experiment showed that, with appropriate training, ratings of TQ made, after viewing the visualizations only, were similar to TQ ratings made after actually watching videos with corresponding video disruptions.
Acknowledgements

I travelled a fair bit while I wrote this thesis and I will not miss having it in my luggage. Thank you to my lab mates: Tiffany Tong, Ryan Kealey, Op and Thomas He for all your help and support in this process. Thank you Mark Chignell, my brilliant supervisor for this opportunity and for the late nights and early mornings. Thank you to my friends Jess, Kelsey, and Merve for keeping me sane. Thank you to my brother for your love, confidence and encouragement. Thank you to my mother and father for always keeping me on the right path. Finally thank you Rob, for walking beside me through it all.
Table of Contents

List of Tables ........................................................................................................ vi
List of Figures .......................................................................................................... vii

Chapter 1: Introduction ............................................................................................. 1
  Motivation ............................................................................................................. 2
  Research Strategy ................................................................................................. 3
  Roadmap of the Thesis ......................................................................................... 4

Chapter 2: Background Literature .......................................................................... 5
  Quality of Experience (QoE) ................................................................................ 5
  QoE Assessment Methods .................................................................................... 6
  Temporal QoE: Momentary, Episodic, and Multi-episodic QoE ......................... 8
  Individual Differences in QoE Assessment: Emotional Intelligence .............. 10
  Video QoE Assessments ..................................................................................... 11
  Video QoE Assessment: Visualizing Versus Viewing Experience .................... 13

Chapter 3: Methodology ......................................................................................... 16
  Participants .......................................................................................................... 16
  Materials ............................................................................................................ 17
    Website ............................................................................................................. 17
    Questionnaires ................................................................................................. 17
    Videos and impairments ................................................................................. 18
    Visualizations ................................................................................................. 20
  Procedure .......................................................................................................... 21
    Prequestionnaire .............................................................................................. 23
List of Tables

Table 1 Description of Video Disruptions .......................................................... 20
Table 2 Correlations in Technical Quality Ratings Between the Sessions ............ 32
Table 3 Cross Tabulation Showing the Relationship Between Training Group and the Four Groups Based on Strength of Correlation Between Session 1 and Session 3
Technical-Quality Ratings .................................................................................. 38
VISUALIZING TECHNICAL QUALITY

List of Figures

Figure 1. Mean opinion score ordinal scale.................................................................8
Figure 2. Final visualization..........................................................................................21
Figure 3. Procedure of Experiment..............................................................................22
Figure 4. A screenshot of the visualization generation software used in Session 2 .......25
Figure 5. Comparison feedback bar in Session 2..........................................................25
Figure 6. Screenshot showing a question being asked in Session 3. .........................26
Figure 7. Feature 1 disagreement ratio calculation......................................................28
Feature 3. UVOFF discrepancy and discrepancy ratio. ............................................30
Figure 9. Changes in technical-quality ratings across the three sessions. ...............33
Figure 10. Impact of training on technical-quality ratings of impairments (left group
of bars represents no training, right group of bars represent with training)...........35
Figure 11. Distribution of correlations between Session 1 and Session 3 technical-
quality ratings (within participants). ........................................................................36
Figure 12. TQ profiles for Session 1 for four groups of participants with different
levels of correlation between Session 1 and Session 2 TQ ratings (with the bars
on the left representing the highest correlations and bars on the right representing
the lowest correlations). .........................................................................................37
Figure 13. Groups based on strength of correlation between session 1 and session 3
technical-quality ratings.........................................................................................40
Chapter 1: Introduction

As online services become more pervasive and complex, customer experience is becoming a key differentiator for business success between service providers (Bharadwan, Varadarajan & Fahy, 1993). In order to avoid high customer churn, increase retention and deal with competitive pressure, information and communications technology (ICT) service providers are seeking to understand the drivers of customer experience, so they can manage the costs of services and meet customer expectations.

The unique mix of increased competition, content-rich resources, and the ‘rise of the consumer’ have resulted in a complicated balancing act with little guidance as to how ICT service providers should keep the costs of sophisticated infrastructures low, while maintaining excellent customer service. The need to find the right balance between provisioning the costs of services and the resulting service quality has led to renewed interest in understanding how users evaluate their experience with a service. Quality of Experience (QoE) in networked communications is an area of research that explores the quality judgments made by a user of a service. It has been defined as “the degree of delight or annoyance of the user of an application or service. It can result from the fulfillment of his or her expectations with respect to the utility or the enjoyment of the application or service in light of the users personality and present state” (Schatz, Hossfeld, Janowski & Egger, 2013).

QoE has been studied in numerous applications including speech communication, text-to-speech synthesis, audiovisual communication, multi-media conferencing, audio transmission, spatial audio rendering, haptics, web browsing, 3D video, and video streaming (Moller & Raake, 2014). For operators and broadcasters, QoE of video
services has become a key service differentiator, along with the number of channels or the content they offer (Staelens, Moens, Broeck, Marien, Vermeulen, Lambert, Ve de Walle, Demeester, 2010). QoE has also become a key factor in routing mechanisms and resource management schemes for network operators and IPTV providers (Balasubramaniam, Mineraud, McDonagh, Perry, Murphy, Donnelly & Botvich, 2011). However, due to the mixture of variables involved in the overall perception of quality, QoE is difficult to measure accurately (Hubbe, Mahfoufi, 2011).

Motivation

Subjective assessment methodologies in QoE research typically have one of, or a combination of, the following three problems (Weiss et al., 2014)

1. QoE is difficult to operationalize. Subjective QoE is a judgement made by a user upon experiencing a service. These judgements can be influenced by a multitude of factors which makes QoE a multi-dimensional construct and inherently difficult to measure.

2. Current subjective-assessment methods are time consuming. To gather QoE assessments, common methods include presenting repetitive stimuli to participants, which reduces the impact of content on quality-assessment judgments by participants, but increases boredom and/or fatigue.

3. The majority of subjective assessment methods are not structured to give information about how judgements change over time with repeated exposure to a service. Subjective assessment methods are primarily focused on momentary, episodic or multi-episodic judgements of QoE.
To address the problems in conducting subjective QoE-assessment, and the relative paucity of video QoE research (Mintauckis, 2010; Weiss et al., 2014), we have developed and tested a new assessment method for QoE in video. The method uses visualizations of impairments and failures to represent disrupted video, with visual cues in the visualizations communicating information about the start time, end time, and duration of the disruption. The visualization method introduced here seeks to employ these images in place of videos, to alleviate long experiment times and to reduce the confounding effects of content bias on QoE judgements.

**Research Strategy**

The overall question addressed in the research reported below is: “Can static, visual representations of a video structure work as a proxy for the viewing experience?” To explore this question, I assessed whether or not participants could learn to recognize images of video disruptions and provide a consistent technical quality (TQ) rating – whether the rating was an image representing the disrupted video, or whether it was based on actually watching the disrupted video. The assessment consisted of an exploratory study where the independent variables/factors were: training/no training condition and video disruption type. The dependent variable was judgment of TQ. Thirty-seven participants were recruited using kijiji.ca and craigslist.com for the experiment, with stimulus materials being presented to participants on a website accessed via a browser. Analysis of the data collected in the experiment addressed two main questions.

1. How well does the profile of TQ across different types of disruption, measured using visualizations, match the corresponding video measured after watching actual videos with corresponding disruptions?
2. To what extent does training to use the visualizations improve the consistency of TQ assessments between ratings based only on visualizations and ratings made after actually watching videos?

**Roadmap of the Thesis**

In the following chapter (two), I review the research literature on QoE and discuss QoE measurement of video. I also review literature on imagery that is potentially relevant to the visualization method introduced in this thesis. Chapter three then reviews the new video QoE assessment method and presents the methodology used in the experiment. Chapter four then presents the experimental results, focusing on how well the pattern of TQ ratings across different video disruptions is preserved when TQ ratings are based on static visualizations of disrupted video, rather than the actual video. Chapter five discusses implications of the results obtained. In the concluding chapter (six) I outline the major contributions that were made, and I also discuss the limitations of the research and prospects for future research on this topic.
Chapter 2: Background Literature

This chapter will begin by introducing the construct of QoE. Methods of QoE will then be discussed, with particular emphasis on QoE assessment for online video. Past work on visualization of experience will also be reviewed briefly.

**Quality of Experience (QoE)**

The most commonly accepted definition of QoE has been provided by a specialized agency of the United Nations that is responsible for issues that concern information and communication technologies, the International Telecommunications Union (ITU). The ITU defines QoE, as: ‘the overall acceptability of an application or service, as perceived by the end user. This includes the complete end-to-end system effects and may be influenced by user expectations and context’ (ITU, 2006). Le Callet, Moller, and Perkins defined QoE as: “the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user’s personality and current state.” This new definition goes beyond a simple acceptance or lack of acceptance of service quality implied in the 2006 ITU definition and addresses the hedonic aspects of quality judgment formation.

To operationalize QoE, the ITU has created specific guidelines regarding test stimuli, ratings scales, room setup and procedure. The QoE evaluation in this standardized environment is well understood and allows for reproducible, reliable results.

Unfortunately, there are also a number of disadvantages to this type of assessment: the kinds of stimuli used are restricted, the number of simultaneous participants is limited, the demography of participants is not reflective of the diversity in
the general public, and the evaluation environment does not reflect the majority of real-life environments where the stimuli are consumed (Hoßfeld, Keimal, Hirth, Gardlo, Habigt, Diepold & Tran-Gia, 2014).

In spite of the strong influence of the ITU in QoE assessment techniques, there have been forays into other kinds of QoE assessment techniques. For instance, researchers have explored the physiological and behavioural aspects of QoE through electroencephalography (Arndt, Antons, Schleicher, Möller, & Curio, 2012), they have changed the testing environment by conducting research in ‘living labs’ (De Moor, Ketyko, Joseph, Deryckere, De Marez, Martens & Verleye, 2010) and have employed crowdsourcing methods to enable larger scale measurements of stimuli (Hobfeld, et al., 2014). In recent years, there has been a change from traditional quality assessment approaches that are network centered (Serral-Garcia et al., 2010), to approaches that are based on the experience of the user.

**QoE Assessment Methods**

Delivering outstanding QoE is critical for video streaming service providers. It is a differentiating factor in reducing churn since users will typically persist with a service as long as their perceived QoE matches their expectations (Hubbe, Mahfoufi, 2011). However, the perception of QoE is inherently subjective and the levels of service expectation vary between users. From the end user’s point of view, the overall QoE of a video streaming service will be influenced by key criteria such as video quality, audio quality, speed of service access, frequency of service interruption, price and pricing model, and after-sales service responsiveness among many other issues. In the context of video transmission, technical quality (TQ), content quality (CQ), and consequently
overall experience (OX) are among the most important influence factors for the service provider (Spachos, Li, Chignell, Zucherman & Jiang, 2015). Due to the mixture of variables involved in the overall perception of quality, it is difficult for operators to measure QoE accurately (Hubbe, Mahfoufi, 2011).

The most commonly used metric to measure subjective video quality is called the Mean Opinion Score (MOS, Figure 1) (Wechsung & De Moor, 2014). It consists of soliciting user opinions to rate video quality after compression and transmission. MOS was originally formulated to evaluate the audio quality of compressed speech in the telecommunications world. It has been used since 1969 (Rothauser, 1969) and continues to be widely used in the quality evaluation of video (Pham-Thi, Hoang-Van & Miyoshi, 2015). MOS yields a rating of between 1 (bad) and 5 (excellent) for perceived service quality (Hubbe, Mahfoufi, 2011). Using MOS requires an extensive battery of tests, in a controlled environment, with a large panel of users.

Although MOS is the current standard metric for subjective QoE, some QoE researchers prefer other measures based on task performance, physiological indicators (e.g., skin conductance as a proxy for cortisol or stress), or user behaviour in general such as cancellation rates or viewing time (Möller & Raake, 2014). However, within the QoE research community, these measures tend to be considered complementary to rather than replacements for MOS rating. Measures beyond MOS and more specifically, beyond ITU recommendations are seen as the exception, rather than the rule (Wechsung & De Moor, 2014).
Temporal QoE: Momentary, Episodic, and Multi-episodic QoE

Due to the increased importance of packet switched networks for media delivery and the resulting temporal fluctuations of media transmission quality, the time-dynamics of experience has become a central topic in audio-visual quality research during the last decade (Hossfeld, Biedermann, Schatz, Platzer, Egger & Fiedler, 2011). Typical fluctuations of media quality take place within time spans (between 15 sec to several minutes) (Gros & Chateau, 2001; Weiss, Meller, Raake, Berger, Ullmann, 2009) that are covered by short-term memory (STM) and working memory, which is based on the interplay between STM and controlled attention, respectively (Baddeley, 2003; Rubin, Hinton, Wenzel, 1999).

Most researchers look at QoE in terms of events that are static and measured at a single point in time. However consumption of services such as online video can occur in episodes that extend over minutes (e.g., a movie trailer) or hours (a film) and videos can be consumed regularly (e.g., when watching a television series over longer periods of time like weeks or months). Three corresponding time spans of user experience are: momentary/instantaneous experience of the service; episodes of usage like a call or video
clip; and multiple episodes that can be recalled as cumulative experience over days, weeks, or months when evaluating a service. Momentary QoE is assessed instantaneously and provides information about the effects of microscopic variation in quality of service (QoS) (e.g. bit rate change). Episodic QoE refers to a quality assessment judgment of a single episode and assessment occurs after the episode. Multi-episodic QoE assessment is particularly relevant to user attitudes toward a service, measuring the accumulation of repeated exposure/experience of the service. Weiss et al. (2014) described multi-episodic QoE as a constant contrast and comparison between users’ internal reference of a service and their individual experiences with the service, which influences their perception of future interactions of the service. Its assessment is not necessarily attached to a single episode, because its scope is the whole service, up to the assessment.

Kahneman (2000) distinguished between momentary experience and memory of experience. For momentary QoE, researchers can observe instantaneous user reaction to changes in QoS. In contrast, the memory-based approach involves retrospective appraisal, or remembered experience, and can apply to episodic and multi-episodic QoE, which is differentiated by the time of assessment and the length of the experience (Weiss et al., 2014). Remembered QoE reflects the users’ integration process as well as the establishment and development of attitudes toward the system/service (Weiss et al., 2014).

Momentary, episodic, and multi-episodic QoE are conceptually distinct, and likely require different assessment measures (Borowiak & Reiter, 2013). Episodic ratings do not easily relate to momentary QoE because the averaged momentary QoE is not an appropriate estimate for the final episode judgment, as it is often too optimistic (Hamberg
VISUALIZING TECHNICAL QUALITY

& de Ridder, 1995). Examples of temporal-assessment approaches include slider-assessment mechanisms (Hamberg & de Ridder, 1995) for momentary QoE, and end-of-episode MOS ratings for episodic QoE. Methods for assessing multi-episodic QoE, have been challenging, but have become especially relevant to service providers because they reflect user-attitude formation toward a service.

Overall, QoS is time varying and presents disparate issues in assessing temporal QoE. Momentary QoE relies on instantaneous assessment, which can often distract users from experiencing the service by forcing them to make continuous judgments using tools such as sliders. General assessment methods in episodic and multi-episodic QoE are not very different, but the cumulative experience measured in multi-episodic QoE raises new issues and challenges for existing methodologies.

Individual Differences in QoE Assessment: Emotional Intelligence

When people make judgments of the technical quality of disrupted video, their judgments may be affected by their emotional responses to impairments and failures or they may assess the technical quality independently of whatever emotions they experience. Emotional Intelligence (EI), as a psychological theory, developed in 1990 by Salovey and Mayer, involves a constellation of constructions related to handling emotions. EI consists of appraising an emotion in the self and others, expression of emotion, regulation of emotion in the self and others, and use of emotion in solving problem (Salovey & Mayer, 1990). In the present context the relevant aspect of emotional intelligence is how aware people are of their emotional responses, how they control them, and how emotional responses affect their ratings of experience. It is
possible that EI may impact how people rate technical quality or the level of reliability of those ratings.

**Video QoE Assessments**

Various studies have been conducted to determine the QoE for online video as a function of QoS: the technical video parameters (resolution, frame rate and codec) and spatial and temporal video artefacts resulting from network imperfections e.g. packet loss, delay, and jitter (Piamrat, Viho, Bonnin, Ksentini, 2009). As of this writing, much of the video content is available via Dynamic Adaptive Streaming over HTTP (DASH), a technique also known as progressive download which is based on HTTP (Hypertext Transfer Protocol) and TCP (Transmission Control Protocol) which assures a reliable, ordered delivery of video packets. Using progressive download, buffering mechanisms and packet retransmissions can avoid the audio-visual distortions due to packet loss and jitter, but may incur re-buffering interruptions and additional start-up delays (Mok, Chan & Chang, 2011).

The assessment of video QoE is a relatively recent phenomenon. For instance, Mintauckis (2010) reported, “research on audio QoE has matured, but research on video QoE is rather new and there is still a lack of a comprehensive metric.” One reason for various approaches to QoE measurement comes from the nature of QoE- it is an experience of a service and experiences are inherently personal, existing only in the mind of an individual (Pine & Gilmore, 1998). This phenomenon can be described as a type of *qualia*, a philosophical term for the irreducible elements of individual instances of subjective, conscious experience where no two people can have the same experience (Pine & Gilmore, 1998; Kanai & Tsuchiya, 2012). As a result, the difficulty with QoE
assessments is that the experience can only exist as viewed from a personal perspective and is only accessible to the responder of the experience (Buck, 1993). In the scientific community, attempts to circumvent the private nature of *qualia* have resulted in the study of ‘qualia behaviours’, defined as “publicly observable responses closely correlated with reports of subjective experience and other emotional responses” (Buck, 1993).

Video QoE assessment falls into one of the three principal types of quality assessment in media applications. These include explicit tests where users evaluate sequences of videos in a controlled setting (i.e., subjective QoE assessment), instrumental-quality assessment algorithms (i.e., objective QoE assessment), or a mix of considerations of context and user behaviour in conjunction with the assessment of technical performance parameters (Garcia et al., 2014).

Most subjective QoE assessments test users’ perceptions of specific degradations due to packet loss or compression. The most common video distortions or degradations include effects such as video jerkiness, dropped video, and freezing or ‘blockiness’ of visuals (Garcia, Argyropolos, Staelens, Naccari, Rios-Quintero & Raake, 2014). Of particular interest to the research in this thesis is subjective QoE in response to freezing and dropped video. As with other QoE assessment methods formulated by the ITU, video subjective QoE assessment is set up using a controlled and standardized method, which can make it difficult to assess external validity of the results. Realistic testing environments need to control the numerous factors that influence a user’s ecosystem. However, because video QoE assessment over IP is relatively new, researchers still need to determine which types of subjective tests and tools are most effective to measure QoE.
In summary, QoE is an emerging field where the definition and operationalization of the term and associated methodologies are still evolving. The multidimensional nature of QoE and its temporal variability make it difficult to measure. Furthermore, the issue of addressing retrospective versus momentary appraisals of QoE also needs to be addressed.

**Video QoE Assessment: Visualizing Versus Viewing Experience**

QoE for online video generally reflects a user’s perception of the technical quality of the video when faced with specific degradations or impairments. Current ITU endorsed methods of obtaining subjective video-over-IP QoE are rigorous in nature and often repetitive in stimuli; they create user exhaustion and boredom, which impacts the ecological validity and reliability of the QoE judgments obtained. The relationship between perception and imagery has been a topic of interest for many years (e.g., Perky, 1910; Scripture, 1896). Wraga and Kosslyn (2003) defined “an image as an internal representation that produces the experience of perception in the absence of the appropriate sensory input”. Finke (1989) defined imagery as “the mental invention or recreation of an experience that in at least some respects resembles the experience of actually perceiving either in conjunction with, or in the absence of, direct sensory stimulation. Mental imagery occurs when “perceptual information is accessed from memory which gives rise to the experience of ‘seeing in the mind’s eye’, ‘hearing with the mind’s ear’ and so on” (Kosslyn, Ganis & Thompson, 2001). Of relevance to the present research, mental images do not have to be formed by recalling of previous experiences, but instead can also be created by combining and modifying stored information in new ways (Kosslyn et al., 2001).
A large body of work is devoted to understanding the processes that may overlap between perception and imagery. Researchers showed that imagery may be able to facilitate perception by priming a common representation (Finke & Kosslyn, 1980; Ishai & Sagi, 1995; Michelon & Koenig, 2002; Pearson, Clifford, & Tong, 2008; Pilotti, Gallo, & Roediger, 2000) and that an interference effect exists between perception and imagery (Craver-Lemley & Arterberry, 2001; Thompson, Hsiao, & Kosslyn, 2011). Researchers have found evidence that imagery and perception draw from the same low-level sensory processes, and that deficits in perception also accompany deficits in the ability to form corresponding images (Bisiach & Luzzatti, 1978; Bourlon, Pradat-Diehl, Duret, & Bartolomeo, 2002; Shuttleworth, Syring, & Allen, 1982; Sparing et al., 2002).

In the field of music, trained musicians are able to experience music, just from reading the score. They imagine the sound of the music through the visual medium of the notes and this skill is taken so seriously in the music world that it is often used as a way to measure musical ability (Schumann & Schumann, 1848/2009).

Mental imagery is an internal representation that produces the experience of perception in the absence of the appropriate sensory input (Wraga & Kosslyn, 2003). Further, mental imagery is cross modal such that, for instance, trained musicians have the ability to ‘hear’ and ‘experience’ music based on visual cues, and mental images do not have to be formed from the recall of previous experiences, but instead can also be created by combining and modifying stored information in new ways. Applying the approach of imagined experience to the problem of QoE assessment in video-over-IP environments, it may be possible to develop static, visual representations of a video experience that could be proxies for actual viewing experience. When such visualizations are appropriately
designed, perhaps with a certain amount of training, users may be able employ mental imagery (perhaps similar in some ways to musical imagery) to experience the technical quality of a video. The experience of technical quality could then be imagined after seeing a visualization of the impairments and failures in the playback of the video. It may also be possible for participants to be trained to associate visual representation of video technical structure with an actual video-viewing experience, to encourage the development of mental imagery and the ‘visualization of experience’ of video technical quality. The next chapter of this thesis will describe such a visualization method.

In summary, QoE is an emerging field where the definition and operationalization of the term and associated methodologies is still evolving. The multi-dimensional nature of QoE and its temporal variability make it difficult to measure. Furthermore, the issue of dealing with retrospective vs. momentary appraisals of QoE also needs to be addressed. Since the method of presenting disrupted videos is effortful and time consuming, it is possible that TQ judgments based on proxies (i.e., visualizations of videos with associated impairments and failures) might be used instead. The next chapter outlines the research methodology that was used in this thesis, including research design, research questions and hypotheses, participant selection and protection of human subjects, and instrumentation.
Chapter 3: Methodology

The challenge for quality assessment is to operationalize QoE in a way that creates reliable and valid measurements. In this chapter, I describe a new assessment method of QoE assessment for video-over IP that uses training, visual proxies of service experience, and mental imagery to circumvent problems of long experiment times, content bias and participant boredom associated with existing methods of QoE assessment. I also report on an initial experiment to determine if the new method of QoE assessment is, in fact, a valid measure of QoE.

Participants

I recruited 37 participants (20 women, 18 men) for the experiment using kijiji.ca, and craigslist.com (online websites for posting and viewing online classified advertisements). I paid participants $20 for participating in the experiment and obtained informed consent from each participant, according to the research protocol that was approved by an ethics review board at the University of Toronto. All participants had normal to corrected vision and reported that English was their first language, or that they had the same fluency in English as a native speaker. Participant age ranged from 18 to 55+ years with the majority of participants in the 25–35-age range \( (N = 17) \) and 18–24 age range \( (N = 13) \). The education level of participants ranged from high school to professional degrees, with the majority of participants having completed some undergraduate education \( (N = 12) \) or having completed an undergraduate degree \( (N = 17) \). All 37 participants were randomly assigned to the two levels of a between-subjects factor, i.e control \( (N = 17) \) or experimental condition \( (N = 20) \). Participants in the experimental condition received training (in the first of three sessions) that showed visualizations of
disruptions to video as they occurred, whereas participants in the control condition did not receive that training. In addition, participants in the experimental condition received feedback as to the correctness of their generated visualizations in Session 2, whereas the participants in the control condition did not receive that feedback.

**Materials**

**Website**

I created a website to present the surveys, videos, and visualizations used in the experiment to participants, with the assistance of Kanmanus Ongvistepaiboon, a graduate student in Computer Science from King Mongkut’s University of Technology Thonburi, Thailand. Mr. Ongvistepaiboon’s participation was part of an ongoing collaborative research program between the School of Information at King Mongkut’s University of Technology, Thonburi and the Interactive Media Lab at the University of Toronto. We used HTML, CSS, JavaScript, PHP, and MySQL, along with the jQuery and CodeIgniter frameworks, to facilitate the development. We used VideoJS as the library for the video player and designed the website collaboratively over the course of several months using wireframes and prototypes to specify interactive elements.

**Questionnaires**

I used three questionnaires in the experiment. The first questionnaire contained demographic questions, the second was an EI questionnaire, and the third was a post-questionnaire that focused on the participants’ experience with the experiment to derive potential gaps for future experimental designs. In addition, a few questions were asked after each video was viewed in the first two sessions of the experiment (Appendix B).
The demographics questionnaire contained six questions on age, gender, and education, as well as on video-viewing preferences. Preference questions included the frequency of video viewing, exposure to video impairments, and attitude toward impairments. These questions were rated on a 5-point frequency-assessment Likert-type scale with 1 corresponding to never and 5 corresponding to always. A full list of the questions can be found in Appendix B.

I administered The Assessing Emotions Scale (AES) (also known as the Emotional Intelligence Scale, the Self-Report Emotional Intelligence Test, or the Schutte Emotional Intelligence Scale), to participants in the experiment. The AES is mainly based on Salovey and Mayer’s (1990) model of EI (as cited in Schutte, Malouff, & Bhullar, 2009). I specifically chose the AES for its brevity, in comparison to other EI scales. The AES is a 33-item self-report inventory that focuses on typical EI. Participants respond to each item by rating themselves on a 5-point Likert-type scale where 1 = strongly disagree and 5 = strongly agree. The scale is scored by reverse coding some items and summing all items. Scores can range anywhere from 33 to 165, where higher scores denote stronger EI.

At the end of the experiment, I presented participants with four open-ended questions about user experience. The post-questionnaire included questions about what they liked and did not like about the experiment, what they found interesting, what they wished could have been different, and how it could have been different.

**Videos and impairments**

Participants saw 30 videos out of a possible 31 videos (with the ordering of the 30 videos randomized for participants) across two sessions of 15 videos each (see Appendix
The video-play times ranged between 60 and 124 seconds. Video-content types varied between short animations and sketches in comedy and drama, as well as movie trailers in the same content categories, namely comedy and drama. I sourced videos from online resources such as Youtube and selected them based on the following screening criteria: no violence, nudity, or offensive language; and between 1 minute and 2 minutes in length. Each video had a resolution of 550 x 300 pixels and an average frame-rate of 25 frames per second (see Appendix C).

Each participant saw each video only once. Two types of video disruption were used in the experiment: integrity impairments and non-retainability failures (see Appendix C). Integrity impairments are exemplified by video buffering where the video freezes for a short period of time, then resumes without losing content. A non-retainability failure is exemplified by a dropped video where the video starts playing but at some point during the playback, drops unexpectedly and cannot be retrieved unless the page is reloaded. The disruptions used in the experiment were derived from past work by Li et al. (2014), (see Table 1).
VISUALIZING TECHNICAL QUALITY

Table 1

Description of Video Disruptions

<table>
<thead>
<tr>
<th>Impairments</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>unimpaired video</td>
</tr>
<tr>
<td>11</td>
<td><strong>One</strong> 10-second integrity impairment that happens @ 10 seconds of video time</td>
</tr>
<tr>
<td>12</td>
<td><strong>Two</strong> 10-second integrity impairment that happen @ 20 seconds of video time and 30 seconds of video time</td>
</tr>
<tr>
<td>13</td>
<td><strong>Three</strong> 10-second integrity impairment that happen @ 10 seconds, 20 seconds, and 40 seconds of video time</td>
</tr>
<tr>
<td>R1–20</td>
<td><strong>One</strong> integrity impairment of 10 seconds that happens @ 10 seconds of video time + a non-retainability failure @ 20 seconds of video content time</td>
</tr>
<tr>
<td>R1_60</td>
<td><strong>One</strong> integrity impairment of 10 seconds that happens @ 10 seconds of video time + a non-retainability failure @ 60 seconds of video content time</td>
</tr>
<tr>
<td>R2_40</td>
<td><strong>Two</strong> 10-second integrity impairments that happen @20 seconds and @ 30 seconds of video time + a non-retainability failure @ 40 seconds of video time</td>
</tr>
<tr>
<td>R2+60</td>
<td><strong>Two</strong> 10-second integrity impairments that happen @ 20 seconds and @ 30 seconds of video time + a non-retainability failure @ 60 seconds of video time</td>
</tr>
</tbody>
</table>

Visualizations

In Session 1 of the experiment, disruptions were shown on a playbar under the video (similar to the kind of playbar seen when playing a Youtube video) but only participants in the experimental condition were shown that playbar. In the second session all the experimental participants (in both the experimental and control conditions) used a tool that let them generate a visualization of the technical quality of each video after they had viewed it. Then in Session 3, only the visualizations were shown and participants were asked to rate the TQ implied by each of the presented visualizations.

The visual format of visualizations was developed over a number of iterations but was ultimately modelled after the visualization of in-video advertisements on YouTube (see Appendix D). I chose this visualization strategy because it was familiar to the majority of the intended users. Most users would have used YouTube and would be familiar with the
progressive playbar and the yellow sections on the playbar that signified the presence of
advertisements. The varying iterations of the visualization design can be seen in the
Appendix E. The final example of the playbar can be seen in Figure 2. Non-retainability
impairments were indicated by a final 'grey' segment that extended to the end of the bar.

Figure 2. Final visualization.

Procedure

I designed the website software for the experiment to handle multiple users at a
time. Experimental sessions were run over two days with six groups of about six people
per group. The procedure for each session was as follows (see Figure 3). A research
assistant welcomed participants to the computer laboratory where the experiment took
place. I pre-set each computer terminal as either a no training or training terminal and
allocated participants randomly to one of the two conditions by assigning them to a
computer terminal. To ensure participants could work independently, there was a
minimum 1.5 metre distance between participants in the experiment. Before starting the
experiment, the research assistant gave participants an overview of what they would be
doing in the experiment, showed participants how to create a visualization using the
visualization software, and answered any questions they had. When there were no more
questions and participants agreed they were comfortable with the tasks, the research
assistant asked them to begin the experiment using the software and desktop computer
provided to them. The first screen the participants engaged with was the consent form.
After reading the form and providing their consent, they began the experiment. The experiment consisted of three sessions, shown schematically in Figure 3.

*Figure 3.* Procedure of Experiment.
VISUALIZING TECHNICAL QUALITY

Prequestionnaire

After reading and signing the consent form, participants in the control and experimental conditions received a set of demographic questions, followed by the AES.

Session 1: ‘Watch and rate’

The experience of participants in Session 1 varied depending on whether they were in the training or no training group. All participants viewed 15 videos one at a time which were either unimpaired, or had one of the disruptions shown in Table 1 (where both the video presented and the disruption presented were randomly generated by the computer software). After viewing each video, participants rated the technical quality (TQ) of the video, the content of the video (CQ), and their overall experience (OQ) using the 5-point MOS rating scale. The only difference between the training group and no training group was that participants in the training group saw the progressive playbar at the bottom of the video as it played, whereas the no training group watched the video without the playbar.

Session 2: ‘Watch, rate, generate’

Session 2 was similar to Session 1 with videos viewed and judgement ratings made on the 5-point MOS rating scale. However, after rating each video, participants were asked to generate a visualization of a playbar with embedded disruptions that matched what they had just seen in the video. To perform this task they used visualization-generation software (see Figure 4) developed especially for this study. Using the tool, they were asked to recreate a playbar to best depict the video they viewed based on the perceived technical quality (TQ) of the video. The task consisted of dragging onscreen sliders appropriately so as to recreate the disruptions that were present
in the video they had just seen. For each impairment or failure they dragged the onscreen slider representing the start of an impairment or failure to the point representing the corresponding starting time of the disruption (the elapsed time within the video when the impairment or failure started). They then dragged the second slider to a point representing the duration of the impairment (instance of freezing). In the case of a non-retainability failure (network failure), the slider shown on the right of Figure 4 was dragged to a point showing the elapsed time in the video where the failure occurred, see Appendix A for interactive elements,

In Session 2, only participants in the training group were shown visual feedback – a comparison between the correct visualization and the participants generated visualization (see Figure 5). Participants then repeated the sequence of watching, rating, and generating 15 times with different videos and impairments. Note that the only difference between the training and no training groups in Session 2 was that the no training group did not get the visual feedback (Figure 5) showing how the visualized playbar should have looked.
Session 3: ‘Rate’

The procedure followed in the third and final session was identical for the training and no training conditions. Participants viewed 28 independent visualizations each, and rated them on their technical quality using the same MOS 1–5 scale (see Figure 6). The visualizations were the same visualizations users had seen in session 1 (for training groups) and that they generated in session 2 (for both training and no training groups). The only difference was that visualizations in session 3 were not accompanied by video (see Figure 6). Visualizations of undisrupted videos (i.e. I0) were not shown, and there were multiple instances shown of each of the different types of disruption used in the earlier sessions (Table 1).
Figure 6. Screenshot showing a question being asked in Session 3.
Chapter 4: Results

Feature Description

In Session 2, users interacted with a software tool to create visualization playbars that indicated the start and end time of each impairment as well as if and when a non-retainability failure occurred. To evaluate the quality of the visualization playbars generated by participants, we developed features to quantify different aspects of accuracy of visualized playbars. The following features were inferred from the user data.

Feature 1: Time Agreement and the Disagree Ratio

The first feature mapped how well the user-generated disruptions (the visualizations, the recreation of their TQ experience of the video) matched the actual video disruptions. The rationale of the features was that I wanted to see what the overall match (agreement) and mismatch (disagreement ratio) of a user-generated disruption (the visualization), and the actual video disruption was in order to gauge the impact of training. This feature measures the degree of overlap between each participant’s visualizations and the actual video and its disruptions. This was done in two ways: through ‘agreement’ which measured the amount of agreement (in seconds) between the users generated visualization, and the actual video disruptions and through a ‘disagree ratio’ which gave a ratio of the amount of disagreement time between the user generated visualization and the actual video total time. The disagreement was recorded as the number of seconds of time, over the length of the video, that the state (playing or interrupted) of the playbar generated by the user did not match the actual playbar corresponding to the video that was watched. I divided this disagreement by the length of
the video to get a ratio indicating how much of the user re-creation ‘disagreed’ with the actual video, which became ‘the disagree ratio’. I calculated the ratio using the following formula involving the total length of elapsed time (including time involving disruptions) and the amount of time for which the user judgment of the state agreed with the actual state: \( \frac{T_{\text{full}} - T_{\text{agree}}}{T_{\text{full}}} \).

A zero-disagreement ratio indicated that the generated playbar was completely accurate. Figure 7 shows an example of how the disagreement ratio was calculated.

In summary, Feature 1 has two parameters: Agree time which describes the amount of time the user re-creation overlapped with the actual video (the larger the number, the better the re-creation) and the Disagree ratio which indicates the ratio of the ‘disagree’ length and the entire length of video (the smaller the ratio, the more accurate the re-creation).

![Figure 7. Feature 1 disagreement ratio calculation.](image)

**Feature 2: Time shift and Agree Time with Time Shift**

Feature 2 is similar to Feature 1 but with a time shift where the first disruption a user notes is shifted forward in time or back in time to match the actual disruption (see Figure 8). In the data it was noted that some users had the right number of disruptions as well as the disruption type but still had a high disagreement ratio despite general agreement with actual disruptions. To account for this and the potential that the issue may
have resulted from the inherent design of the visualization tool, we wanted to see how much of a shift was required for users and if the shift would change the overall agreement. Feature 2 had two parameters: ‘time shift’, the amount of shift (in seconds) needed to align the first disruption of the users visualization to match the actual disruption and ‘agree with time shift’, the amount of time (in seconds) when the user recreation of disruptions overlapped with actual video disruptions. I calculated the Feature 2 disagreement ratio after the user-generated playbar was slid right or left (forward or backward in time) so that the first video disruption time matched between the user-generated and actual disruptions. A positive number meant the user-regenerated video was shifted forward in time and a negative number meant the user-regenerated video was shifted back in time.

Thus two parameters describe Feature 2: Agree time with time shift and time shift. Agree time with time shift describes the amount of time for which the user-generated visualization (after being shifted to match the actual video disruption) where the larger the number, the better the re-creation. Time shift indicates the amount of shift (in seconds) needed to align the first disruption of the user-regenerated video with the actual video disruption where a positive number meant the user-regenerated video was shifted forward in time.
Feature 3. UVOFF discrepancy and discrepancy ratio.

Feature 3 compared the number of disruptions generated by the user with the actual number of disruptions. This feature had one parameter that was measured in two ways, the first was the ‘UVOFF discrepancy’ where the number of impairments the user over or underestimated was expressed as a positive or negative (e.g. if a user failed to create one disruption, the OFF discrepancy would have had a value of -1, if they added in an extra 2 disruptions, it would be +2). The second measurement method was the ‘Discrepancy ratio’ where the number of features overestimated or underestimated was expressed as a percentage (ranging from -100% to +100%). The number of disruptions a participant generated was divided by the number of actual disruptions.

Feature 4: Off-Time Ratio

Feature 4 measured whether a user over- or under-estimated the duration of the disruption (in seconds). The first step in creating the feature was to sum all the user-generated disruption times, and the actual disruptions times. The feature was then calculated as the ratio of total user disruption time to total actual disruption time. The resulting quantity was referred to as the ‘OFF-time ratio’. An OFF-time ratio greater than
one suggests that the user overestimated the duration of the interruptions while a ratio of less than one indicates underestimation.

**Analysis**

The main focus of the analysis was to address the question of how well people can judge technical quality (TQ), based only on visualizations of video interruptions. To answer this question, I compared the profile of TQ scores across the different types of disruptions in Session 3 (visualizations only) with corresponding profiles for Session 1 (watching video with or without a playbar showing the disruptions), and Session 2 (making TQ judgments prior to generating visualizations). Prior to comparing the TQ profiles across sessions I also examined the correlation of TQ scores (within individuals) between Session 1 (actual videos viewed) and Session 3 (visualizations only). Table 2 shows these correlation results. The (Pearson) correlation in TQ judgments between Session 1 and Session 2 was .732 (i.e., 53.6% of the shared variance). In contrast, the correlations of Session 3 with Sessions 1 and 2, respectively, were .587 and .606, or roughly 35% of the shared variance. Thus, by this measure, TQ judgments captured a significant portion of the TQ judgments (around two thirds).
Table 2

Correlations in Technical Quality Ratings Between the Sessions

<table>
<thead>
<tr>
<th></th>
<th>Correlations</th>
<th>TQ_mean.1</th>
<th>TQ_mean.2</th>
<th>TQ_mean.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>TQ_mean.1</td>
<td>Pearson correlation</td>
<td>1</td>
<td>.732</td>
<td>.587</td>
</tr>
<tr>
<td></td>
<td>Sig. (1-tailed)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>239</td>
<td>207</td>
<td>239</td>
</tr>
<tr>
<td>TQ_mean.2</td>
<td>Pearson correlation</td>
<td>.732</td>
<td>1</td>
<td>.606</td>
</tr>
<tr>
<td></td>
<td>Sig. (1-tailed)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>207</td>
<td>232</td>
<td>232</td>
</tr>
<tr>
<td>TQ_mean.3</td>
<td>Pearson correlation</td>
<td>.587</td>
<td>.606</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sig. (1-tailed)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>239</td>
<td>232</td>
<td>266</td>
</tr>
</tbody>
</table>

Note. TQ = technical quality.

Figure 9 shows the TQ ratings across the three sessions. The general pattern of results was preserved, with failures (non-retainability) judged more harshly than (integrity) impairments. The pattern of TQ mean ratings for failures in Session 3 seemed to match the corresponding pattern for Session 1 better than the equivalent pattern for Session 2. Although the pattern of the Session 3 data tends to match Session 1, the mean ratings tend to be slightly lower (by a small fraction of a rating point) with the exception of the r2_40 failure, which had roughly the same mean rating in Sessions 1 and 3 (see Figure 9).
Figure 9. Changes in technical-quality ratings across the three sessions.

How Well Does Visualized TQ Agree with Experienced TQ?

In the introductory chapter to this thesis I posed the following question: How well does the profile of TQ across different types of disruption, measured using visualizations, match the corresponding video measured after watching actual videos with corresponding disruptions? In order to answer this question I conducted a repeated-measures ANOVA with two factors: session (Session 1 vs. Session 3) and disruption (the I1, I2, and I3 impairments and the four non-retainability failures, making seven levels in total). For the impairment main effect and the impairment by session interaction, significant evidence emerged of sphericity, assessed by Mauchly’s test. Thus, I used the Greenhouse-Geisser
correction to the degrees of freedom in assessing those effects. A statistically significant effect emerged for session \((F[1, 7.57] = 4.54, p < .05)\) with TQ ratings in Session 3 tending to be lower, as can be seen in Figure 9. Also, a significant main effect of impairment emerged, as expected \((F[3.29, 121.6] = 96.91, p < .001)\). Crucially, there was no sign of any interaction that arose between session and impairment \((F = 1.47, \text{NS})\). Thus, the TQ profile across the seven types of disruption, when participants judged visualizations, was similar to the corresponding TQ profile when participants judged actual videos with the corresponding disruptions. I conducted a similar repeated-measures ANOVA comparing Sessions 2 and 3. Again, a significant main effect emerged for session \((p < .05)\) and disruption \((p < .001)\) but there was no significant session by impairment interaction \((F < 1)\).

**Does Training Improve the Validity of Visualized TQ Judgments?**

The second research question posed in the introductory chapter was: To what extent does training to use the visualizations improve the consistency of TQ assessments between ratings based only on visualizations and ratings made after actually watching videos. In order to answer this question I conducted a mixed ANOVA on the Session 3 data to see if training had an impact on the TQ ratings of disruptions. The between-subjects factor was the condition (training or no training) and the within-subjects factor was the disruption. A significant interaction emerged between disruption and training \((F[2.79, 90.36] = 3.34, p < .05)\). Figure 10 shows this interaction, where the two groups of bars represent Session 3 TQ ratings with no training on the left, and with training on the right. In each group, the seven bars correspond to the seven video disruption conditions used in the study. As can be seen in Figure 10, without training, failures were judged less harshly, with their ratings
roughly equivalent to the ratings for I3. However, for those who received training, a clear separation arose between types of disruptions (impairments and failures), and the general patterns of mean TQ ratings across the seven levels of disruptions (impairments and failures) was quite similar to the pattern obtained in Session 1. This pattern of results shows that training was beneficial in helping people make realistic TQ ratings (i.e., ratings that matched how they would have rated actual videos with those disruptions) on the basis of the visualizations only.

*Figure 10. Impact of training on technical-quality ratings of impairments (left group of bars represents no training, right group of bars represent with training).*
I carried out an additional analysis to see if people who were more consistent in their ratings between sessions 1 and 3 would also provide TQ profiles that were more closely related to the mean TQ profile obtained across participants when viewing the actual videos with disruptions in Session 1.

*Figure 1*. Distribution of correlations between Session 1 and Session 3 technical-quality ratings (within participants).

Figure 11 shows the distribution of correlations between Session 1 and Session 3 mean TQ ratings within participants. Ten participants have a correlation above .9 (Group 1) and six participants have correlations between .8 and .9 (Group 2). Nine people have correlations between .6 and .8 (Group 3). Ten participants had correlations between zero
and .6 (Group 4). The remaining two people had negative correlations between their Session 1 and Session 3 TQ ratings.

Figure 12 shows the TQ ratings for Session 1, blocked by the four levels of correlation between Session 1 and Session 3 TQ ratings referred to in the preceding paragraph.

![TQ profiles for Session 1 for four groups of participants with different levels of correlation between Session 1 and Session 2 TQ ratings](image)

*Figure 12. TQ profiles for Session 1 for four groups of participants with different levels of correlation between Session 1 and Session 2 TQ ratings (with the bars on the left representing the highest correlations and bars on the right representing the lowest correlations).

People in Group 1 (correlations above .9) and Group 2 (correlations between .8 and .9) strongly separated impairments from failures, but showed no apparent difference
in TQ ratings between I2 and I3. Group 3 (correlations greater than or equal to .6) showed less separation between I1 and I2 and no distinction between I3 and the failure. For Group 4 (correlations less than .6), the distinctions between the seven disruption types were less clear with smaller differences between all the disruption types. Overall, the results shown in Figure 11 do not show a strong relationship between the reliability of participants between sessions 1 and 3 and the validity of their TQ assessments as assessed by the degree of differentiation between impairments and failures. However, people with low reliability between sessions 1 and 3 (i.e. with correlations below .6) did show weaker distinctions between the disruptions in their ratings. This suggests that people whose TQ ratings change markedly when they use visualizations are less likely to provide reliable TQ judgments when rating actual videos.

In the mixed ANOVA reported earlier, training was beneficial in generating TQ ratings of visualizations of disruptions that had a similar profile (across the disruptions) to corresponding ratings of impairments and failures that were actually viewed. I checked to see if this training benefit could be attributable to higher correlations between Session 1 and Session 3 TQ ratings. I grouped participants according to the size of the correlation in the TQ ratings between Session 1 and Session 3 using the four groups defined above. Table 3 shows the number of training and no training participants in each range. Training group participants were spread across all four of the correlation ranges. Thus, there is little evidence that training improved the correlation between session 1 and 3 TQ ratings.

Table 3

Cross Tabulation Showing the Relationship Between Training Group and the Four
VISUALIZING TECHNICAL QUALITY

Groups Based on Strength of Correlation Between Session 1 and Session 3 Technical-Quality Ratings

<table>
<thead>
<tr>
<th>CorrelationGroup</th>
<th>TrainingGroup</th>
<th>Count</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No training</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Training</td>
<td>20</td>
<td>37</td>
</tr>
</tbody>
</table>

In a final analysis, I considered whether training impacted the measures of visualization-generation performance. I conducted separate *t*-tests on each of the features that assess quality of generated visualizations (explained above) with training vs. no training as the grouping factor. Four evaluative measures showed significant differences with training (*p < .01*), whereas the other two did not significantly differ with training (*p > .1*) (see Figure 13).
The four significant measures were:

- AGREE (Feature1)
- DISAGREE RATIO (Feature 1)
- AGREE with shifting (Feature 2)
- UVOFF DISCREPANCY (Feature 3)

Agree (Feature 1) and agree with shifting both are higher in the training group, and the discrepancy ratio and UV off discrepancy are lower in the training group. This suggests that users in the training group (versus the no training group) are better at generating the right number of disruptions, with the start times and lengths of durations being more
VISUALIZING TECHNICAL QUALITY

aligned to the real disruptions. In the following chapter I will discuss the practical significance of the results reported in this chapter.
Chapter 5: Discussion

This chapter discusses the main findings of the experiment and their implications for the use of visualizations of video disruptions on a playbar as a way of collecting TQ ratings. The discussion is organized around the following questions.

1. Can people generate visualizations of disruptions? Is the concept of using a playbar in this way even reasonable? The answer to this question is clearly yes. All 37 participants in this study were able to generate visualizations and no issues emerged concerning the face validity of the task. This readiness to accept visualizations may be because people are used to seeing advertisements marked in yellow on YouTube playbars, where the advertisements function somewhat analogously to impairments.

2. Are the patterns of impairments made when viewing visualizations similar to those made after viewing actual video with impairments? Repeated-measures ANOVAs compared Session 1 to Session 3, and Session 2 to Session 3, with impairments as a second factor. For both these analyses, no sign of an interaction arose between the session and impairment effect. In other words, in characterizing video disruptions, no evidence emerged that TQ ratings made using visualizations were any different from TQ ratings made after watching corresponding videos. This statistical result is supported by visual inspection of the corresponding bar charts, where the pattern of TQ ratings across impairment types is remarkably similar for Session 1 and Session 3, although the Session 3 (visualization only) ratings are a bit lower than the corresponding Session 1 ratings in general.
3. Is training people to use visualizations for TQ rating beneficial? The answer to this question was yes. The pattern of ratings of impairments in Session 3 was significantly different for people who had been trained, and their results better matched the Session 1 mean results, where people viewed the actual video. The effect of training was relatively independent of the correlation between Sessions 1 and 3. Stability in the overall pattern of ratings across the impairments is a better indicator of the adequacy of the visualization method than is the reliability of correlations of TQ ratings of visualizations versus actual videos, within individuals.

4. Would training help people generate visualizations better? One possible pathway for using visualizations is to first learn how to generate them. By making the connection between the actual disrupted video and how the corresponding visualization would look, it should then be possible to reverse the process and imagine the experience of viewing the impaired video based on viewing the visualization. A series of $t$-tests identified four (of six) measures of visualization-generation performance that improved significantly with training. Therefore, in addition to improving TQ ratings based on viewing visualizations, training also helped people generate visualizations that better matched what they actually saw/experienced.

5. Are TQ ratings made with visualizations reliable, in the sense that they are correlated with ratings made after watching actual videos with disruptions? The answer to this question is that this correlation was statistically significant, and in terms of variance accounted for, it is about two thirds the size of the
correlation between repeated TQ ratings (between Sessions 1 and 2, where the disrupted videos continue to be viewed). However, reliability may not be a good measure of how well visualizations worked. Although training was effective, there was no strong tendency for training to lead to higher measured reliability (correlations between Session 1 and Session 3 in this case). Also, no marked tendency emerged for people with the highest correlations between Sessions 1 and 3 to have better-looking patterns of TQ rating than people with more moderate (but not low) correlations, although people with very low correlations between Sessions 1 and 3 did show lower sensitivity to TQ differences due to video disruptions.

In summary, the visualization method for judging TQ associated with disrupted video appears to work, provided people are trained. In the experiment reported here, a relatively simple regimen of showing disruptions on a playbar as videos played, and then providing feedback when people generated playbar visualizations based on the experience of watching disrupted video, proved to be sufficient. Thus, not only can training enable people to provide reasonable TQ ratings when viewing visualizations (only) of disrupted videos, but the amount of training necessary to achieve this ability is relatively modest.
Chapter 6: Conclusions

**Contributions**

The research presented here demonstrates the reliability of a new proposed QoE-assessment tool. The contributions made in this research are as follows.

1. A novel method for showing disruption video as visualizations was developed.
2. An experimental methodology was designed, and implemented on a website, for assessing the validity of visualized video for the purpose of measuring video QoE.
3. The obtained research results showed that visualized TQ ratings are a good approximation of the TQ ratings that are provided in response to viewing actual video (for the types of video disruption considered in this research).
4. The research results also showed that training people to understand visualizations and generate them is effective in improving the validity of visualized TQ ratings.

In summary, this work has added to the growing body of work that looks past traditional QoE assessment measures. Users benefited from training and were able to complete tasks reliably and consistently across sessions. However, as with any new tool, limitations exist.

**Limitations**

Based on the present results it is not clear that the TQ estimates made by the trained participants in response to viewing visualizations were any less valid than the TQ estimates after viewing the corresponding videos. However, only a relatively small
sample of disruptions were tested and it is possible that more subtle distinctions in TQ may be more difficult to differentiate using visualizations. However, for the disruption examples considered in the present study, there was no evidence of a loss of accuracy that accompanied the greater efficiency provided by the use of visualizations. As a rough estimate of the gain in efficiency provided by the use of visualizations in the present study, participants in Session 1 (judging TQ based on viewing actual video) took roughly 17 minutes to judge 15 videos. In contrast it took between 5 and 10 minutes to judge the 28 visualizations in Session 3. It should also be noted that I used fairly short video clips that are around a minute to a minute and a half in length and that the relative efficiency of the visualization method will increase as the length of the video that has to be viewed increases.

Only a limited set of video disruptions were used in this study and the differences between the types of disruption were relatively coarse-grained. Thus, it remains to be seen how fine the distinctions between different types of disruption can be when using visualizations. In addition the videos used in this experiment were relatively short and it remains to be determined how well the present results will scale to longer videos. Further, the experiment was run in a computer lab, which lacked ecological validity as participants normally experience impairments and failures in their regular routines and familiar environments. It is possible that users would have different reactions to non-retainability impairments if they were able to react naturally (e.g. go get a snack or have the option open a different browser). Further, only desktop computers were used in this experiment. It may also be the case that users would react much differently to integrity
and non-retainability impairments if they were on a mobile phone, where the environment around them impinges on the viewing experience.

Participants were also collectively tested which may have impacted their concentration, comfort and subsequent judgement of QoE. Finally, due to time constraints, analyses of individual differences in EI and video watching habits were not factored into the analysis of the data and their impact on the visualization tool.

One further speculation about the benefit of visualization that has yet to be demonstrated is that visualization may allow people to assess the impact of multiple video viewing sessions over extended periods of time. However, new training methods and new experimental evaluations would be needed to demonstrate the feasibility of using visualization to measure cumulative video viewing experiment over days or even weeks.

**Future Research**

Despite the limitations of this study, next steps in this line of research could further iterate the design of the visualizations and the presentation of experimental and control conditions to participants. In regard to the design of the visualizations it may be the case that participants who have better time perception may perform better at this type of assessment tool and, thus, the next set of tests should also assess time perception ability.

More research into the visualization generation tool could also be carried out, including iterative design with user testing. Further, individual differences in users could be examined more closely in conjunction with the visualization tool to determine if the tool caters to certain individuals better than others and why. For instance, the impact of
individual differences in time perspective, affective forecasting, emotional regulation, patience, and video-watching habits could be examined in relation to how people use the visualization tool.

It may also be the case that experiences on different screen size, devices (i.e. tablet, mobile phone, TV or laptop) affect QoE judgements. Future studies could also study visualizations in the context of services that a user repeatedly engages with over time, such as chat, gaming or video conferencing.

Finally, taking it a step further, the visualization tool could be reimagined to be a tool for both episodic and multi-episodic QoE. For instance, researchers could present multiple visualizations (similar to those in Session 3) simultaneously to represent multiple episodes and ask the user to imagine they are experiencing them and forecast what their end decision would be if they experienced the service represented in the visualizations. This application of the visualization tool could potentially provide a useful observational tool to address the question of how attitude and behaviour toward a service is shaped, based on a user’s longer-term experience with a service.
References


VISUALIZING TECHNICAL QUALITY


VISUALIZING TECHNICAL QUALITY


VISUALIZING TECHNICAL QUALITY

Schatz, R., Hossfeld, T., Janowski, L., & Egger, S. (2013). From packets to people: Quality of experience as a new measurement challenge. In E. Bliersack, C. Callegari, & M. Matijasevic (Eds.), *Data traffic monitoring and analysis* (pp. 219–263). Berlin, Germany: Springer-Verlag


doi:10.1002/0470018860.s00549
Appendix A: Interactive Elements
VISUALIZING TECHNICAL QUALITY

Visualizing Technical Quality
Please use these visualization tools to recreate the technical quality of the video you just watched. Try to make your visualization as accurate as you possibly can.

Video Freezing
A 6-second impairment starts from 10 seconds

The video froze at...
The freezing lasted for this many seconds...

Network Failure
The network error occurred at...

Apply this network failure

Apply this freezing impairment
VISUALIZING TECHNICAL QUALITY

Visualizing Technical Quality
Please use these visualization tools to recreate the technical quality of the video you just watched. Try to make your visualization as accurate as you possibly can.

Video Freezing
A 6-second impairment starts from 10 seconds | remove
The video froze at...

Network Failure
The network error occurred at...

Visualizing Technical Quality
Please use these visualization tools to recreate the technical quality of the video you just watched. Try to make your visualization as accurate as you possibly can.

Video Freezing
A 6-second impairment starts from 10 seconds | remove
The video froze at...

Network Failure
A retainability failure starts from 39 seconds | remove
The network error occurred at...

Visualizing Technical Quality: Playbar Comparison
True visualization of technical quality
Your visualization of technical quality
Appendix B: Questions for Participants

Directions: Each of the following items asks you about your emotions or reactions associated with emotions. After deciding whether a statement is generally true for you, use the 5-point scale to respond to the statement. Please circle the “1” if you strongly disagree that this is like you, the “2” if you somewhat disagree that this is like you, “3” if you neither agree nor disagree that this is like you, the “4” if you somewhat agree that this is like you, and the “5” if you strongly agree that this is like you.

There are no right or wrong answers. Please give the response that best describes you.

1 = strongly disagree  
2 = somewhat disagree  
3 = neither agree nor disagree  
4 = somewhat agree  
5 = strongly agree

1. I know when to speak about my personal problems to others.  
2. When I am faced with obstacles, I remember times I faced similar obstacles and overcame them.  
3. I expect that I will do well on most things I try.  
4. Other people find it easy to confide in me.  
5. I find it hard to understand the non-verbal messages of other people.  
6. Some of the major events of my life have led me to re-evaluate what is important and not important.  
7. When my mood changes, I see new possibilities.  
8. Emotions are one of the things that make my life worth living.  
9. I am aware of my emotions as I experience them.  
10. I expect good things to happen.  
11. I like to share my emotions with others.  
12. When I experience a positive emotion, I know how to make it last.  
13. I arrange events others enjoy.  
14. I seek out activities that make me happy.  
15. I am aware of the non-verbal messages I send to others.
<table>
<thead>
<tr>
<th>Question</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>16. I present myself in a way that makes a good impression on others.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>17. When I am in a positive mood, solving problems is easy for me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>18. By looking at their facial expressions, I recognize the emotions people are experiencing.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>19. I know why my emotions change.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>20. When I am in a positive mood, I am able to come up with new ideas.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>21. I have control over my emotions.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>22. I easily recognize my emotions as I experience them.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>23. I motivate myself by imagining a good outcome to tasks I take on.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>24. I compliment others when they have done something well.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>25. I am aware of the non-verbal messages other people send.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>26. When another person tells me about an important event in his or her life, I almost feel as though I experienced this event myself.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>27. When I feel a change in emotions, I tend to come up with new ideas.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>28. When I am faced with a challenge, I give up because I believe I will fail.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>29. I know what other people are feeling just by looking at them.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>30. I help other people feel better when they are down.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>31. I use good moods to help myself keep trying in the face of obstacles.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>32. I can tell how people are feeling by listening to the tone of their voice.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>33. It is difficult for me to understand why people feel the way they do.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Session#2 Confidence Rating

How confident are you that your visualization accurately depicts the technical quality of the video?

- Extremely
- Very Moderately
- Slightly
- Not at all

Next

Session#3 Rating

Your overall evaluation of the technical quality in the video is:

- Excellent
- Good
- Fair
- Poor
- Bad
## Appendix C: List of 31 Videos

<table>
<thead>
<tr>
<th>Video Name</th>
<th>Resolution</th>
<th>Frame Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACloudyLesson</td>
<td>640 x 364</td>
<td>29.97</td>
</tr>
<tr>
<td>AlabamaMoon</td>
<td>640 x 272</td>
<td>24</td>
</tr>
<tr>
<td>Arbitrage</td>
<td>640 x 364</td>
<td>24</td>
</tr>
<tr>
<td>ASeriousMan</td>
<td>640 x 338</td>
<td>23.97</td>
</tr>
<tr>
<td>BeautifulBoy</td>
<td>640 x 364</td>
<td>24</td>
</tr>
<tr>
<td>BrokenThings</td>
<td>640 x 364</td>
<td>25</td>
</tr>
<tr>
<td>Committed</td>
<td>480 x 360</td>
<td>23.97</td>
</tr>
<tr>
<td>CzarnyPunkt</td>
<td>640 x 364</td>
<td>25</td>
</tr>
<tr>
<td>Fugu</td>
<td>640 x 360</td>
<td>24</td>
</tr>
<tr>
<td>GraceOfMonaco</td>
<td>640 x 360</td>
<td>25</td>
</tr>
<tr>
<td>Gumball</td>
<td>480 x 360</td>
<td>25</td>
</tr>
<tr>
<td>HappyEverAfters</td>
<td>640 x 360</td>
<td>25</td>
</tr>
<tr>
<td>HelpWanted</td>
<td>640 x 364</td>
<td>30</td>
</tr>
<tr>
<td>KissingBooth</td>
<td>640 x 352</td>
<td>23.97</td>
</tr>
<tr>
<td>LifeInADay</td>
<td>640 x 364</td>
<td>30</td>
</tr>
<tr>
<td>LightsOut</td>
<td>640 x 360</td>
<td>25</td>
</tr>
<tr>
<td>LionsForLambs</td>
<td>480 x 272</td>
<td>23.97</td>
</tr>
<tr>
<td>Madison</td>
<td>640 x 344</td>
<td>23.80</td>
</tr>
<tr>
<td>Meek</td>
<td>640 x 360</td>
<td>29.97</td>
</tr>
<tr>
<td>OpalDream</td>
<td>640 x 360</td>
<td>25</td>
</tr>
<tr>
<td>PixarShortFilms...</td>
<td>640 x 364</td>
<td>23.97</td>
</tr>
<tr>
<td>PoirotElephants...</td>
<td>640 x 350</td>
<td>25</td>
</tr>
<tr>
<td>Prom</td>
<td>640 x 364</td>
<td>30</td>
</tr>
<tr>
<td>RobotsAndButterfly</td>
<td>640 x 360</td>
<td>24</td>
</tr>
<tr>
<td>Saawariya</td>
<td>640 x 272</td>
<td>23.97</td>
</tr>
<tr>
<td>TheGreatBeauty</td>
<td>640 x 364</td>
<td>25</td>
</tr>
<tr>
<td>TheOpenRoad</td>
<td>640 x 360</td>
<td>25</td>
</tr>
<tr>
<td>Time</td>
<td>640 x 360</td>
<td>23.97</td>
</tr>
<tr>
<td>WarOfTheButtons</td>
<td>640 x 360</td>
<td>25</td>
</tr>
<tr>
<td>WendyAndLucy</td>
<td>640 x 364</td>
<td>24</td>
</tr>
<tr>
<td>WinWin</td>
<td>640 x 360</td>
<td>24.38</td>
</tr>
</tbody>
</table>
Appendix D: In-Video Advertisements on YouTube
Appendix E: Iterations of the Visualization Design