IMPACT OF MOBILE DEVICES AND USAGE LOCATION ON PERCEIVED MULTIMEDIA QUALITY

Andrew Catellier, Margaret Pinson, William Ingram and Arthur Webster

{acatellier, margaret, bing, webster}@its.bldrdoc.gov

ABSTRACT

We explore the quality impact when audiovisual content is delivered to different mobile devices. Subjects were shown the same sequences on five different mobile devices and a broadcast quality television. Factors influencing quality ratings include video resolution, viewing distance, and monitor size. Analysis shows how subjects' perception of multimedia quality differs when content is viewed on different mobile devices. In addition, quality ratings from laboratory and simulated living room sessions were statistically equivalent.

Index Terms— multimedia, audiovisual, subjective testing, coder/decoders, mobile, standards

1. INTRODUCTION

Multimedia-capable mobile devices are proliferating. These devices, together with improving and more ubiquitous network connections, are influencing the ways in which users expect to interact with mobile devices. The heterogeneity of network conditions, use cases, and devices presents new questions and therefore new challenges in the realm of subjective testing, quality of service, and quality of experience measurement. For example: is an audiovisual sequence that is of sufficient quality on a ten-inch screen still of sufficient quality when the ten-inch device is made to display the sequence on a large-screened television? Alternatively, if a user switches from watching content on a large-screened television to watching content on a 3.5 inch device, how much can the transmission rate be lowered while still providing acceptable quality? What if the mobile device user is watching the sequence while traveling on public transportation?

The growing multitude of new multimedia-capable devices makes it increasingly difficult for any one research laboratory to be able to conduct research representative of the entire mobile device market. Subjective testing is famously expensive, and the current ITU Recommendations were not designed for newer display technologies or high-quality mobile displays [1]–[3].

To ensure repeatable results, ITU Recommendations specify procedures and laboratory testing environments for multimedia subjective tests, but these are very different from where mobile devices are typically used. Can we expect these laboratory environments to capture the user's experience?

Users of mobile devices may derive some incremental enjoyment or increased utility when watching audiovisual content on the bus to work or while waiting in a grocery line. The distance between a user's eyes and the screen of the mobile device cannot be assumed to be constant. Pervasive background noise often exists in a mobile device's most likely use location.

The authors are with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, Colorado 80305.

It is imperative to understand how all of these variables affect a user's perception of audiovisual content. It would be valuable to understand how tests using mobile devices and non-standard testing locations relate to standardized testing procedures, and therefore to data available from years and years of subjective testing. In order to achieve this goal, rudimentary experiments must explore the mechanics of testing mobile devices in multiple testing environments, including less-than-satisfactory and changing environmental conditions.

We describe an exploratory experiment using five mobile devices in two testing environments. The mobile devices used in the experiment ranged from a smartphone to a large laptop. They were tested using audiovisual sequences delivered via a web interface in two environments: an ITU-T Recommendation P.911 compliant laboratory [3] and a simulated living room. The same set of audiovisual sequences was also evaluated on the broadcast quality monitor in the ITU-T P.911 test chamber for comparison purposes.

Section 2 discusses the relevant existing literature in the field. Section 3 lays out this test's procedures and details. Finally, the results of the experiment and the conclusion are reported in Sections 4 and 5 respectively.

2. RELATED WORK

Researchers in related fields have conducted studies into the interactions between mobile devices and audiovisual quality. Gulliver *et al.* used a head-mounted display as well as a computer and a PDA in a study investigating mobile devices [4]. This study found that subjects rated audiovisual quality lower when viewing content on a head-mounted display. The authors suggest that distortion caused by the head-mounted display (which was designed to simulate a 52-inch screen, and incidentally had a larger field of view than the PDA) may have been the reason for the lower ratings.

In a later study, Gulliver and Ghinea used a PDA and a head-mounted video display system along with headphones to perform another exploration of mobile multimedia quality [5]. Findings in this work suggest that device type had no significant effect on subjects' level of enjoyment, but did have a significant effect on the perceived audiovisual quality.

Chen *et al.* used a web interface and an MPEG1 plugin for Microsoft's Internet Explorer to measure various aspects of multimedia perception and how cognitive styles influence perception [6]. In [7], Song *et al.* conducted a test that emulated delivering audiovisual content to a mobile phone. They noted that higher spatial resolution may make viewers more sensitive to quality degradations (in the form of framerate/bandwidth changes). However, this test was conducted with SIF (Source Input Format) and QCIF (Quarter Common Intermediate Format) size video. The Video Quality Experts Group (VQEG) conducted an experiment using video resolutions used on mobile devices, but this study did not use mobile

devices [8]. Agboma and Liotta conducted a multimedia test on a mobile phone, a PDA and a laptop, described in [9]. They varied codec rates to create eight quality levels and found that mobile device users will deem multimedia quality acceptable differently on different devices. Research has been done [10] using mobile devices in "living labs"—letting experiment participants use a mobile device during their daily routine. Delivery of multimedia content to a mobile phone using Digital Video Broadcasting for Handhelds (DVB-H) was tested in the context of a train station, a bus, and a café in [11]. During this study, Jumisko-Pyykkö and Hannuksela found that participants were more willing to accept highly degraded audiovisual sequences when using the mobile phone in a location other than a laboratory. Kaikkonen *et al.* found that some problems encountered when using a mobile device are more tolerable when the user is not in the laboratory [12].

In summary, there have been many studies looking into the interaction of mobile devices and audiovisual quality. We learned that results have been compared among devices and that device type can affect quality ratings. Mobile device usage context and field of view can also affect quality ratings. However, our survey does not reveal a study that compares results from mobile audiovisual quality tests to the body of audiovisual studies that comply with standing ITU Recommendations. Further, investigations into how subjective testing scores are affected by environments that aren't standards-compliant are in their infancy. The remainder of this paper describes a stepping stone on the path to understanding the relationships among the huge body of audiovisual experiment results and past and future mobile audiovisual experiment results.

3. EXPERIMENT DESIGN

3.1. Experiment Overview

Our goal was to understand how the perceived quality of audiovisual sequences differed among many devices. We showed the same audiovisual sequences on a variety of mobile devices and a broadcast quality monitor. Each mobile device was tested in two different environments: an ITU-T Rec. P.911 compliant laboratory and a simulated living room.

We chose eight source sequences and three video quality levels: "excellent," "fair" and "bad." To reduce the size of the experiment, the original audio was used for all sequences. The original video was used for the "excellent" quality level. The "fair" and "bad" video quality levels were coding distortions. The bitrate for each sequence was separately adjusted to appear "fair" and "bad" to us when played on the broadcast monitor. This allowed us to ensure that all three quality levels were represented evenly. The lowest video quality level was chosen to be *at most* "bad" quality—such that a sequence wasn't likely to receive a score higher than "bad"—on the broadcast-quality monitor.

3.2. Test Hardware and Software

During this experiment, six different displays were tested:[1] an iPod Touch 3rd generation (**iPT**), an iPhone 4 (**iP4**), a first-generation iPad (**iPad**), a Sony laptop with a 15 inch screen (**S15**), a Dell with a 17 inch screen (**D17**) and a broadcast-quality LCD television and speakers (**BM**).

The iPod Touch has processing capability and a display that is roughly equivalent to current typical smartphones. The iPhone 4's display is the same size as the iPod Touch's, but the iPhone 4's resolution is double that of the iPod Touch. Testing these two devices simultaneously may allow interesting conclusions to be drawn. The iPad is a competitor in a new market for thin, touch-screen tablet computers, and was the only such device available for sale when equipment for this experiment was procured.

The Sony and Dell were chosen to represent different laptop markets. The Sony has a 15 inch display and is blue and slim. It represents a consumer-grade laptop. The large, gray Dell has a high-resolution 17 inch display and is meant to be desktop replacement or heavy-duty gaming laptop. Pertinent display characteristics of all devices are shown in Table 1. For each device, the pixel dimensions listed are the screen size (top) and video resolution (bottom). In order to listen to audio, subjects were allowed to use either over-the-ear headphones or earbuds.

The broadcast quality television was a TV-Logic LVM-460WD professional grade 46 inch liquid crystal display (LCD) monitor. Audio was delivered using NHT speakers placed on the floor on either side of the display; left and right speakers (model A-20) and left and right subwoofers (model B-20). The sequences were played to the broadcast quality TV and speakers using a Blu-Ray disc player and presented with constant timing.

The subjective test interface was an interactive web application. This web interface was designed to simulate a video viewing website where users can also rate the video. A Mac Pro (early 2009 or 4,1) with 16 GB memory, two 2.26 GHz quad-core processors, and a SATA 2.0 hard drive was used to deliver the processed audiovisual sequences and record votes using modern web technologies. The server was programmed to offer a unique sequence randomization to each subject and record their votes.

The interface of this experiment was driven by web technologies. An Apache web server (version 2.2.17) [13] interacted with a PostgreSQL database (version 9.0.3) [14] and a PHP module (version 5.3.5) [15] in order to serve the content. Hypertext markup language (HTML) including cascading style sheets (CSS) [16] and JavaScript [17] were used to detect which device was accessing the server and render the interface elements accordingly.

The iPhone, iPod Touch, and iPad all had iOS version 4.3 installed, and used Mobile Safari to access and display the test. The two laptops both had a standard Windows 7 64-bit installation (build 7600, Home Premium on the Sony and Professional on the Dell), and used Google Chrome (version 11) to access and display the test. The iPod Touch, iPhone 4 and iPad all have hardware H.264 decoders that are capable of decoding video exceeding their display resolution. All of the mobile devices were tested to ensure smooth video playback.

3.3. Audiovisual Sequence Preparation

We examined dozens of high-definition (HD) sequence, all 1920×1080 pixels, and either 24 or 30 frames per second. We selected nine sequences that exhibited a wide variety of visual and audio effects. Eight sequences were chosen to use in the test, and the remaining sequence was chosen for use during a practice session. Screen captures from the eight audiovisual sequences used in the bulk test are shown in Figure 1. Most of these sequences can be downloaded royalty-free for research and development purposes from the Consumer Digital Video Library [18].

[1] Certain commercial equipment, software, and services are identified in this report to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is necessarily the best available for this purpose.

Table 1. Device Display Characteristics

	width	height	pixel density
iPod Touch	7.5 cm	5 cm	64 p/cm
	480 px	320 px	
video res.	480 px	272 px	
iPhone 4	7.5 cm	5 cm	128 p/cm
	960 px	640 px	
video res.	960 px	544 px	
iPad	19.6 cm	14.7 cm	52.2 p/cm
	1024 px	768 px	
video res.	1024 px	576 px	
Sony	31 cm	17.4 cm	44.1 p/cm
	1366 px	768 px	
video res.	1280 px	720 px	
Dell	36.6 cm	23 cm	52.5 p/cm
	1920 px	1200 px	
video res.	1920 px	1080 px	
Broadcast Monitor	101.8 cm	57.3 cm	18.9 p/cm
	1920 px	1080 px	
video res.	1920 px	1080 px	

Fig. 1. Selected frames from each audiovisual sequence.

Each sequence was imported into the Adobe Premiere CS5 video-editing suite and then edited down to ten seconds. Audio levels were normalized and a short audio fade was placed at the beginning and end of each sequence. In order to create visual impairments, the audiovisual sequences were then compressed using Main Concept's H.264/AVC Pro (version 1.6.34425.0). Each sequence was encoded twice: once to create the medium-quality version ("fair"), and once to create the low-quality version ("bad"). Depending on the scene, bitrates used for medium-quality encoding ranged from 750 kbps to 2000 kbps. Bitrates used for low-quality encoding ranged from 100 kbps to 250 kbps. The AAC audio codec was used to encode the audio of all stereo sequences at a bitrate of 224 kbps regardless of video encoding quality. Adobe Premiere was then used to create uncompressed versions of all medium and low-quality audiovisual sequences. Deinterlacing and frame conversions, if necessary, were done using the AviSynth (version 2.5.8.5) software package. This process resulted in a total of 27 uncompressed HD sequences (i.e., nine original-quality, nine medium-quality, and nine low-quality). These sequences were arranged in two unique orders with no repeated scenes and were burned directly to Blu-Ray for display on the broadcast-quality monitor.

The following describes the process used to convert the uncompressed HD sequences into a format suitable for delivery over IP networks and playback in modern web browsers. A shell script was created to automate this process. The Handbrake [19] command line interface tool (version 0.9.4, using the x264 encoder [20]) was used to convert the sequences into `.mp4` format. Default conversion options were used with the following exceptions: the output video width was varied to match the native resolution of each device, the H.264 profile was specified [21] on a per device basis, and the `q` (constant quality) factor was specified. The optimum `q` was chosen to maximize encoding quality while keeping file sizes small enough to ensure flawless playback on a private wireless network. When the output video width was less than 1920 pixels, the encoder downsampled the video to the resolution noted in Table 1. The percent downsampled and specific encoder settings are shown in Table 2. Resulting bitrates for each quality level and each device are shown in Table 3. Using these settings, a loop in the shell script encoded each of the 27 sequences into a format suitable for playback on each of five mobile devices, resulting in a total of $5 \times 27 = 135$ files. The files were stored on the web server.

The aspect ratio of the audiovisual sequences was preserved in each case. Therefore, the audiovisual sequences did not use all of the available screen real estate on each device. We chose this trade-off rather than cropping the audiovisual sequences uniquely for each device and therefore potentially hiding a different set of impairments for each device.

3.4. Test Procedure

After completing vision screening, subjects were shown an example high-quality audiovisual sequence on each device. Subjects were allowed to handle all devices and were encouraged to use each device's interface to play or pause the sequence. In order to familiarize the subject with the test interface, the experiment administrator then began a practice session for the first mobile device. A practice session was administered once for each device. The practice session consisted of viewing and rating three versions of the practice sequence (high, medium, and low quality). The remaining 24 sequences were

Table 2. x264 Encoder Settings

	width	profile	q	% downsampled
iPT	480	3.0	0.6	75%
iP4	960	3.1	0.55	50%
iPad	1024	3.1	0.6	53%
S15	1280	3.1	0.55	33%
D17	1920	3.1	0.7	0%

Table 3. Average Video Bitrate (in kilobits per second)

	iPT	**iP4**	**iPad**	**S15**	**D17**
high	446	1,023	1,745	2,433	29,859
medium	417	919	1,464	2,170	11,411
low	404	857	1,365	2,025	9,289

rated during the main part of the experiment.

The interface looked consistent among all the mobile devices. At the start of each session, the subject was prompted for a user number and location (lab or living room). Upon pressing the "submit" button, a "play" icon appeared (indicating a video could be started). After the subject touched (on the touchscreen devices) or clicked (using the laptops) the "play" icon, an audiovisual sequence was displayed in the center of the screen. Once the sequence finished playing, five radio buttons appeared under the video display spanning its width. The radio buttons were labeled "excellent," "good," "fair," "poor," "bad" from left to right (these labels correspond the numerals 5, 4, 3, 2 and 1 shown in the figures and tables in Section 4). Once a subject touched a radio button or its associated text, the radio button indicated the choice and a button labeled "vote" appeared. Subjects were allowed to replay the video. After the "vote" button was touched or clicked, it and the radio buttons disappeared and a "play" icon appeared, indicating that a new sequence was ready to be played.

Because the video on the professional monitor was driven by a standalone Blu-Ray player, its voting interface was a little different. The subject would watch the video on the monitor and then vote using the iPod Touch or iPhone 4. The voting interface was similar to that described above, but instead of a video display, text indicating which sequence would be receiving the vote was displayed. A JavaScript timer prevented the voting interface from being displayed before the sequence on the screen had finished playing. Before each sequence, text displaying the number of the upcoming sequence was displayed for five seconds. After each sequence, text reading "Vote for clip X" (where X was the corresponding sequence number) was displayed for five seconds.

Each subject evaluated all processed sequences on each mobile device in each of two testing environments: a standards-compliant sound isolation chamber (S) and a simulated living room (L). The broadcast-quality monitor was only tested in the standards-compliant environment. The use of a sound isolation chamber is expected practice in audiovisual quality research today. Our chamber is compliant to ITU-T Rec. P.911 and ITU-R Rec. BT.500. The simulated living room contained a comfortable chair, a table and an office chair. Indirect sunlight shone through a window. Though generally quiet, background noise was occasionally audible (e.g., a car driving past, sprinklers). The living room was decorated with a photograph of brightly colored leaves, a lamp, and plush toys. For both environments, the subject was alone in the room during the test.

In the standards-compliant environment, subjects were read directions from a script that simply explained how to do the experiment. In the living room, the script also encouraged them to adjust the lighting and move from the comfortable chair to the office chair and table if they desired. The distance between the subject and the five mobile devices was not fixed in either environment, and users were allowed to adjust their viewing distance during the test. However, the distance between the subject and the professional monitor was fixed at approximately three times the picture height.

Thirty subjects completed the experiment. Each was screened for visual acuity and color blindness. One subject was color blind, but the subject's results were not found to be outliers. Thirteen of the subjects started in the simulated living room and 17 started in the sound isolation chamber.

The test consisted of 11 sessions in total. In each session, the subject rated all 24 sequences on one device in one environment. All sessions in one environment were completed before moving to the other environment. The environments, devices, and sequences were randomized differently for each subject. Thus, each subject had unique randomizations for: the order of environments, the order of devices (within each of the two environments) and the order of the 24 sequences (within each of the 11 sessions).

Within each session, a given sequence was not played twice in a row. Subjects were allowed to take breaks between sessions after rating all the sequences on each device.

Sessions lasted between five and ten minutes each, and the entire test took between two and a half and three hours per user. Subjects were paid for four hours of labor and encouraged to take breaks. Subjects participated in the test one at a time due to wireless bandwidth restrictions.

4. RESULTS

A total of 30 (21 male, 9 female) subjects participated in the experiment. Each subject rated all 24 audiovisual sequences on each of the five devices in both environments and on the professional monitor in the lab environment for a total of $(30 \times 24 \times 5) \times 2 + (30 \times 24) = 7920$ votes. Since subjects were allowed to choose a comfortable viewing distance for each device, viewing distance varied among viewers. A measurement between the device and the subject's forehead was taken at the end of each session. These measurements, along with known device specifications were used to calculate the data in Table 4. Average angle subtended per pixel (β_{pix}) and average angle subtended by the display (β_w horizontal and β_h vertical) were calculated using $\beta = 2 \arctan\left(\frac{\frac{w}{2}}{l}\right)$, where β represents angle subtended, w is either pixel or display width or height, and l is the distance from the display to the viewer's eye. β_{pix} is reported in arc-minutes, β_w and β_h are reported in degrees, and average viewing distance (d) is listed in screen heights. Each measure was calculated for each subject and then each measure was averaged for inclusion in the table. Also listed in the table are mean opinion scores (MOS) for medium-quality sequences averaged over all users and all scenes for each device in the standards-based environment (S^m_{MOS}) and the simulated living room (L^m_{MOS}).

It is interesting to note that the iPhone 4's high-resolution display did not seem to have an effect on either S^m_{MOS} or L^m_{MOS} when compared to the iPod Touch. A Student's t-test (1% uncertainty) confirms this ($p = 0.28$). Because the iPhone 4 and the iPod Touch have the same size screen and statistically similar MOS, we can say that, in the context of this experiment, delivering iPod Touch resolu-

Table 2. x264 Encoder Settings

	width	profile	q	% downsampled
iPT	480	3.0	0.6	75%
iP4	960	3.1	0.55	50%
iPad	1024	3.1	0.6	53%
S15	1280	3.1	0.55	33%
D17	1920	3.1	0.7	0%

Table 3. Average Video Bitrate (in kilobits per second)

	iPT	**iP4**	**iPad**	**S15**	**D17**
high	446	1,023	1,745	2,433	29,859
medium	417	919	1,464	2,170	11,411
low	404	857	1,365	2,025	9,289

rated during the main part of the experiment.

The interface looked consistent among all the mobile devices. At the start of each session, the subject was prompted for a user number and location (lab or living room). Upon pressing the "submit" button, a "play" icon appeared (indicating a video could be started). After the subject touched (on the touchscreen devices) or clicked (using the laptops) the "play" icon, an audiovisual sequence was displayed in the center of the screen. Once the sequence finished playing, five radio buttons appeared under the video display spanning its width. The radio buttons were labeled "excellent," "good," "fair," "poor," "bad" from left to right (these labels correspond the numerals 5, 4, 3, 2 and 1 shown in the figures and tables in Section 4). Once a subject touched a radio button or its associated text, the radio button indicated the choice and a button labeled "vote" appeared. Subjects were allowed to replay the video. After the "vote" button was touched or clicked, it and the radio buttons disappeared and a "play" icon appeared, indicating that a new sequence was ready to be played.

Because the video on the professional monitor was driven by a standalone Blu-Ray player, its voting interface was a little different. The subject would watch the video on the monitor and then vote using the iPod Touch or iPhone 4. The voting interface was similar to that described above, but instead of a video display, text indicating which sequence would be receiving the vote was displayed. A JavaScript timer prevented the voting interface from being displayed before the sequence on the screen had finished playing. Before each sequence, text displaying the number of the upcoming sequence was displayed for five seconds. After each sequence, text reading "Vote for clip X" (where X was the corresponding sequence number) was displayed for five seconds.

Each subject evaluated all processed sequences on each mobile device in each of two testing environments: a standards-compliant sound isolation chamber (S) and a simulated living room (L). The broadcast-quality monitor was only tested in the standards-compliant environment. The use of a sound isolation chamber is expected practice in audiovisual quality research today. Our chamber is compliant to ITU-T Rec. P.911 and ITU-R Rec. BT.500. The simulated living room contained a comfortable chair, a table and an office chair. Indirect sunlight shone through a window. Though generally quiet, background noise was occasionally audible (e.g., a car driving past, sprinklers). The living room was decorated with a photograph of brightly colored leaves, a lamp, and plush toys. For both environments, the subject was alone in the room during the test.

In the standards-compliant environment, subjects were read directions from a script that simply explained how to do the experiment. In the living room, the script also encouraged them to adjust the lighting and move from the comfortable chair to the office chair and table if they desired. The distance between the subject and the five mobile devices was not fixed in either environment, and users were allowed to adjust their viewing distance during the test. However, the distance between the subject and the professional monitor was fixed at approximately three times the picture height.

Thirty subjects completed the experiment. Each was screened for visual acuity and color blindness. One subject was color blind, but the subject's results were not found to be outliers. Thirteen of the subjects started in the simulated living room and 17 started in the sound isolation chamber.

The test consisted of 11 sessions in total. In each session, the subject rated all 24 sequences on one device in one environment. All sessions in one environment were completed before moving to the other environment. The environments, devices, and sequences were randomized differently for each subject. Thus, each subject had unique randomizations for: the order of environments, the order of devices (within each of the two environments) and the order of the 24 sequences (within each of the 11 sessions).

Within each session, a given sequence was not played twice in a row. Subjects were allowed to take breaks between sessions after rating all the sequences on each device.

Sessions lasted between five and ten minutes each, and the entire test took between two and a half and three hours per user. Subjects were paid for four hours of labor and encouraged to take breaks. Subjects participated in the test one at a time due to wireless bandwidth restrictions.

4. RESULTS

A total of 30 (21 male, 9 female) subjects participated in the experiment. Each subject rated all 24 audiovisual sequences on each of the five devices in both environments and on the professional monitor in the lab environment for a total of $(30 \times 24 \times 5) \times 2 + (30 \times 24) = 7920$ votes. Since subjects were allowed to choose a comfortable viewing distance for each device, viewing distance varied among viewers. A measurement between the device and the subject's forehead was taken at the end of each session. These measurements, along with known device specifications were used to calculate the data in Table 4. Average angle subtended per pixel (β_{pix}) and average angle subtended by the display (β_w horizontal and β_h vertical) were calculated using $\beta = 2 \arctan\left(\frac{\frac{w}{2}}{l}\right)$, where β represents angle subtended, w is either pixel or display width or height, and l is the distance from the display to the viewer's eye. β_{pix} is reported in arc-minutes, β_w and β_h are reported in degrees, and average viewing distance (d) is listed in screen heights. Each measure was calculated for each subject and then each measure was averaged for inclusion in the table. Also listed in the table are mean opinion scores (MOS) for medium-quality sequences averaged over all users and all scenes for each device in the standards-based environment (S^m_{MOS}) and the simulated living room (L^m_{MOS}).

It is interesting to note that the iPhone 4's high-resolution display did not seem to have an effect on either S^m_{MOS} or L^m_{MOS} when compared to the iPod Touch. A Student's t-test (1% uncertainty) confirms this ($p = 0.28$). Because the iPhone 4 and the iPod Touch have the same size screen and statistically similar MOS, we can say that, in the context of this experiment, delivering iPod Touch resolu-

Table 4. Subject distance and angle info

	β_{pix}	d	β_w	β_h	S^m_{MOS}	L^m_{MOS}
iPT	$1.42'$	7.9	$11.3°$	$7.6°$	3.9 ± 0.13	4.0 ± 0.13
iP4	$0.71'$	8.1	$11.4°$	$7.6°$	4.0 ± 0.1	3.9 ± 0.13
iPad	$1.48'$	3.1	$24.8°$	$18.8°$	3.6 ± 0.14	3.6 ± 0.18
S15	$1.32'$	3.5	$29.3°$	$16.8°$	3.2 ± 0.14	3.1 ± 0.15
D17	$1.06'$	2.8	$32.7°$	$20.9°$	2.8 ± 0.19	2.8 ± 0.18
BM	$0.99'$	3.2	$30.9°$	$17.7°$	3.0 ± 0.15	n/a

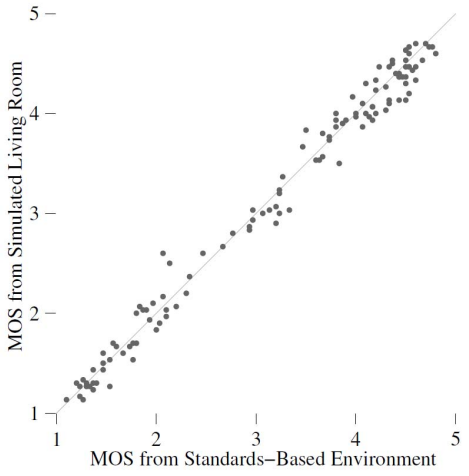

Fig. 2. MOS value for each sequence from the standards-based environment shown on the x-axis, MOS value for each sequence from the simulated living room shown on the y-axis. Correlation is 0.992.

tion sequences to the iPhone 4 would have no negative consequence but allows for a significant bandwidth savings.

As device size increases, the visual angle subtended by the device increases, and except for the iPhone 4, the angle subtended by a single pixel decreases. We suspect that the higher MOS achieved by the small devices is partially due to the subsampling performed on the audiovisual sequences.

Device size also seems to affect the results. The video shown on the iPhone 4 has nearly as many pixels as the video shown on the iPad, but S^m_{MOS} and L^m_{MOS} on the iPad are noticeably lower than on the iPhone 4. However, both the iPod Touch and the iPhone 4 were viewed at an average of about eight display heights. If it were comfortable (or possible) for a subject to view either of these smaller devices at a distance of three picture heights (similar to the average viewing distance of the other devices), it might be possible to perceive distortions more readily.

Figure 2 shows the MOS for each processed sequence in a scatter plot, with the MOS from the standards-based environment on the x-axis and the MOS from the simulated living room on the y-axis. The data has a correlation of 0.992, indicating that a less-controlled environment did not have a significant effect on MOS. Figure 3 shows the MOS for each processed sequence in a scatter plot, but excludes all votes from each subject's second environment. Its correlation is 0.985, so we can say that a subject's familiarity with the test material may only contribute slightly to the similarity of MOS between the two environments.

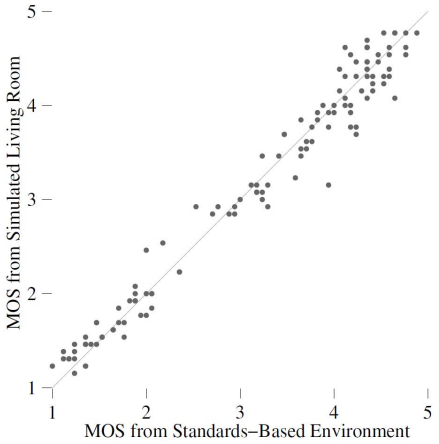

Fig. 3. MOS value for each sequence from the standards-based environment shown on the x-axis, MOS value for each sequence from the simulated living room shown on the y-axis, *excluding votes from a subject's second test environment.* Correlation is 0.985.

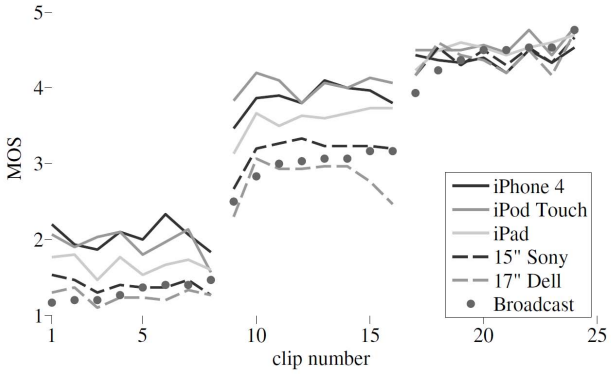

Fig. 4. MOS per sequence for each device from the standards-based environment.

Figure 4 shows the MOS for each sequence based on votes from the standards-based environment. The sequence order was determined by rank-ordering the MOS of each sequence as viewed on the broadcast-quality monitor. The figure shows that the eight high-quality sequences achieved nearly the same MOS on each device. However, the range of MOS measured for the eight medium-quality sequences approaches 1.5 points with the smallest devices receiving scores in the region of "good" and the largest devices receiving scores in the "fair" region. Low-quality sequences exhibit similar behavior, with small screens receiving scores in the "poor" region and larger screens receiving scores closer to "bad." Figure 5 presents the average MOS as a function of the device's video resolution. The black line and confidence intervals (CI) are S_{MOS}; and the grey line and CI are L_{MOS}. Figure 5 shows that for high-quality sequences, all devices received statistically similar MOS. It also shows that the smartphone/tablet form factors achieve statistically different scores than the devices with larger screens for medium and low-quality sequences.

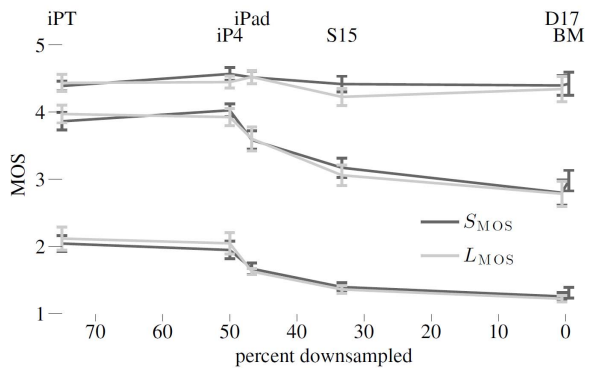

Fig. 5. MOS per device for high, low and medium quality. The x-axis shows how much downsampling was required to get the video down to each device's native resolution.

5. CONCLUSIONS

The purpose of this test was to investigate the effect on multimedia quality when audiovisual sequences are transcoded or subsampled to suit a number of mobile devices. We found that high-quality sequences are subjectively rated just as highly on small mobile devices as they are on large high-definition televisions. We can also conclude that very significant bandwidth savings are possible if a content server is aware of what kind of device is requesting audiovisual content and sends content suited for the device.

Since S_{MOS} and L_{MOS} are not significantly different, we can begin to compare these results to some existing research that did not take place in ITU-compliant testing environments. We can therefore draw upon a rich body of existing research in the field of human-computer interaction.

The infrastructure developed—a web application accessible by most devices with a web browser—to conduct this experiment represents a unique way forward in the area of mobile device subjective testing. The software could be expanded, hardened for security and distributed widely through the Internet, allowing for data collection on a massive scale. The experiment's exploration into non-standard testing environments lays the groundwork for understanding how to interpret data collected from mobile devices wherever they may be in use. Additionally, the results show that a standards-compliant laboratory is not required to achieve stable results, thus lowering subjective testing's barrier to entry.

6. REFERENCES

[1] C. Fenimore, V. Baroncini, T. Oelbaum, and T.K. Tan, "Subjective testing methodology in MPEG video verification," in *SPIE Conference on Applications of Digital Image Processing XXVII*, 2004.

[2] ITU-R Recommendation BT.500-12, "Methodology for the subjective assessment of the quality of television pictures," Geneva, 2009.

[3] ITU-T Recommendation P.911, "Subjective audiovisual quality assessment methods for multimedia applications," Geneva, 1998.

[4] SR Gulliver, T. Serif, and G. Ghinea, "Pervasive and standalone computing: the perceptual effects of variable multimedia quality," *International Journal of Human-Computer Studies*, vol. 60, no. 5, pp. 640–665, 2004.

[5] S.R. Gulliver and G. Ghinea, "Defining user perception of distributed multimedia quality," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 2, no. 4, pp. 241–257, 2006.

[6] S.Y. Chen, G. Ghinea, and R.D. Macredie, "A cognitive approach to user perception of multimedia quality: An empirical investigation," *International Journal of Human-Computer Studies*, vol. 64, no. 12, pp. 1200–1213, 2006.

[7] S. Song, Y. Won, and I. Song, "Empirical study of user perception behavior for mobile streaming," in *Proceedings of the Tenth ACM International Conference on Multimedia*. ACM, 2002, pp. 327–330.

[8] VQEG, "Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase I," Sep. 2008.

[9] F. Agboma and A. Liotta, "User centric assessment of mobile contents delivery," in *Proceedings of 4th International Conference on Advances in Mobile Computing and Multimedia*, 2006.

[10] T. De Pessemier, K. De Moor, A.J. Verdejo, D. Van Deursen, W. Joseph, L. De Marez, L. Martens, and R. Van de Walle, "Exploring the acceptability of the audiovisual quality for a mobile video session based on objectively measured parameters," in *Third International Workshop on Quality of Multimedia Experience (QoMEX)*, Sep. 2011, pp. 125–130.

[11] S. Jumisko-Pyykkö and M.M. Hannuksela, "Does context matter in quality evaluation of mobile television?," in *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM, 2008, pp. 63–72.

[12] T. Kallio and A. Kaikkonen, "Usability testing of mobile applications: A comparison between laboratory and field testing," *Journal of Usability Studies*, vol. 1, pp. 4–16, 2005.

[13] The Apache Software Foundation, "The Apache HTTP Server Project," Website, Available http://httpd.apache.org/.

[14] PostgreSQL Global Development Group, "PostgreSQL," Website, Available http://www.postgresql.org/.

[15] The PHP Group, "PHP: Hypertext Preprocessor," Website, Available http://www.php.net.

[16] W3C, "HTML5, A vocabulary and associated APIs for HTML and XHTML," Website, Available http://www.w3.org/TR/html5/.

[17] Ecma International, "Standard ECMA-262 ECMAScript Language Specification," Website, Available http://www.ecma-international.org/publications/standards/Ecma-262.htm.

[18] Margaret Pinson, Stephen Wolf, Neha Tripathi, and Chin Koh, "The consumer digital video library," in *Fifth International Workshop on Video Processing and Quality Metrics (VPQM)*, Scottsdale, Arizona, USA, 2010.

[19] HandBrake Project, "HandBrake," Website, Available http://handbrake.fr/downloads.php.

[20] VideoLAN Organization, "x264," Website, Available http://www.videolan.org/developers/x264.html.

[21] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560 –576, Jul. 2003.