

Multimedia Synchronization Study

Carolyn Ford
Mark A. McFarland
William Ingram
Scott Hanes
Margaret Pinson
Arthur Webster
Kelsey Anderson



technical memorandum

Multimedia Synchronization Study

Carolyn Ford
Mark A. McFarland
William Ingram
Scott Hanes
Margaret Pinson
Arthur Webster
Kelsey Anderson



U.S. DEPARTMENT OF COMMERCE
Gary Locke, Secretary

Lawrence E. Strickling, Assistant Secretary
for Communications and Information

October 2009

DISCLAIMER

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

CONTENTS

	Page
FIGURES	vi
TABLES.....	vi
ABBREVIATIONS/ACRONYMS.....	vii
1 INTRODUCTION.....	1
2 SUBJECTIVE TEST DESIGN	3
2.1 Source Sequences	3
2.2 Subjective Test Methodology.....	5
3 DATA ANALYSIS.....	6
4 FUTURE WORK.....	7
5 REFERENCES.....	8
APPENDIX A: VIEWER INSTRUCTIONS.....	9
APPENDIX B: TEST SOFTWARE DESCRIPTION	10
APPENDIX C. AUDIO/VIDEO CALIBRATION.....	13

FIGURES

	Page
Figure 1. The SG 9 multimedia quality full diagram.....	2
Figure 2. The sequence of creating testing samples.....	5
Figure 3. MOS as a function of audio offset.....	6
Figure B-1. Subjective test control program interface used to specify type of experiment	10
Figure B-2. Subjective test control program interface used to specify test sequences and practice session	11
Figure B-3. The rating screen showing the MOS scale.....	11
Figure C-1. The equipment setup for the data validation test.....	12

TABLES

	Page
Table 1. Descriptions of Source Scenes.....	3

ABBREVIATIONS/ACRONYMS

ACR	Absolute Category Rating
CIF	Common Intermediate Format (352 pixels by 288 lines)
DOC	Department of Commerce
FPS	frames per second
GUI	graphical user interface
HDTV	High-Definition Television
ITS	Institute for Telecommunication Sciences
ITU-T	International Telecommunications Union, Telecommunications Sector
MOS	mean opinion score
NTIA	National Telecommunications and Information Administration
NTSC	National Television System Committee
PC	personal computer

MULTIMEDIA SYNCHRONIZATION STUDY

Carolyn Ford, Mark A. McFarland, William Ingram, Scott Hanes, Margaret Pinson, Arthur Webster, and Kelsey Anderson¹

ITS is conducting a series of studies to quantify the effects of the separate audio and video compression qualities, and the differential delay in their synchronization, to the perceived aesthetic quality of a multimedia signal. The experiment described in this report was specifically designed to study the effects of the differential delay.

Key words: audio delay; audio offset; audio-video synchronization; differential delay; multimedia quality; subjective testing

1 INTRODUCTION

Much work has been done to characterize human subjective quality of video and audio independent of each other [1][2]. Less work has been done to answer the question of how audio quality and video quality combine to form a person's opinion of an audio-video sequence when viewed and heard simultaneously. We are interested in discovering mathematical functions that describe audio-visual (or "multimedia") quality². Specifically, we would like to explore whether audio-visual quality can be described as a function of the audio quality and video quality measured separately. The multimedia synchronization study was one step towards this goal.

Figure 1 shows how the International Telecommunication Union, Telecommunications Sector (ITU-T), Study Group 9, via Recommendation J.148 [3], envisions that audio quality (Aq) differential delay, and video quality (Vq) can be combined to estimate three quantities:

1. Audio quality in the presence of video, labeled $Aq(Vq)$.
2. Overall multimedia quality, or the quality of the audio-video sequence taken as a whole.
3. Video quality in the presence of audio, labeled $Vq(Aq)$.

¹The authors are with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, CO 80305.

² The term "quality" in this context refers to the subjective aesthetic impression reported by a human viewer, and not to the intelligibility (or intelligence value) of the multimedia sample for human or automatic recognition.

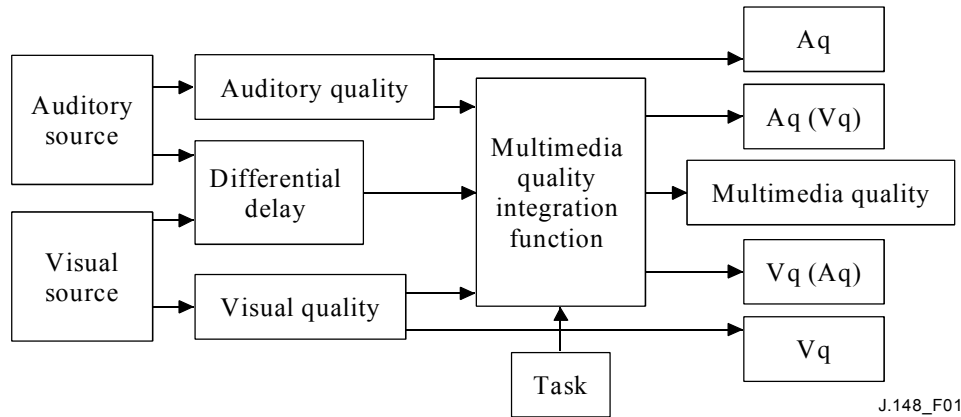


Figure 1. The SG 9 multimedia quality full diagram.

Recommendation J.148 defines the remaining components of the model as:

- The core inputs to the multimedia model are the audio model (e.g., ITU-T Rec. P.862 [4])
- The core inputs to the video/composite image model (e.g., ITU-T Rec. J.144 [1])
- A measure of differential auditory-visual delay (ITU-R Rec. BT.1359-1 [5])

The task of the multimedia model is to integrate together the qualities from the audio and video/composite image models. The output from the multimedia model is a prediction of multimedia quality representative of human perception.

The Institute for Telecommunication Sciences (ITS) is conducting a series of studies that focus on the prediction of overall multimedia quality given the core inputs. The task for these studies is the aesthetic impression of multimedia signals, of varying video resolutions that have been processed by the system under test. For these studies the system under test is transmission effects (compression and desynchronization).

A previous study³, was conducted to quantify the effects of the separate audio and video quality levels on the perceived quality of the multimedia signal in Common Intermediate Format (CIF), when the audio and video are perfectly synchronized. In terms of Figure 1, this is the case in which the differential delay is zero, and the audio quality and video quality are independently varied.

The goal of the multimedia synchronization study was to quantify the effect of audio and video differential delay on the multimedia quality when both the audio and video qualities are constant, with no additional transmission errors beyond the differential delay.




³ McFarland, M., et. al., "Relating Audio and Video Quality, Using CIF Video," ITS Technical Memo (Draft)

2 SUBJECTIVE TEST DESIGN

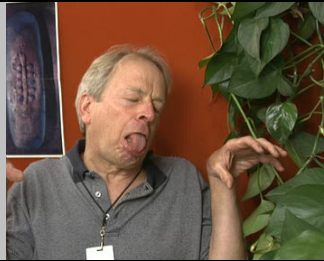





2.1 Source Sequences

The test subjects were shown CIF resolution video combined with 16 kHz mono audio. The nine source audio-visual sequences used in this experiment are described in Table ⁴. Each of these sequences has obvious audio-video synchronization clues (e.g., lip synchronization, percussion sounds from a strike). These audio-visual sequences were originally filmed in either National Television System Committee (NTSC) format (720 pixels by 486 lines, 29.97 fps) or High-Definition Television (HDTV) 1080i 60Hz (1920 pixels by 1080 lines, 29.97fps). All source sequences were edited to 14 seconds duration, and digitally converted to CIF resolution, with no additional compression applied.

Table 1. Descriptions of Source Scenes

File Name	Video Description	Audio Description	Example Video Content
Boxingbags	Man trains with boxing bags	No voice, percussive hits	
Boystalk	Two boys talking in a dynamic lighting environment	Two young boys' voices	
Cartalk1	Boy tells a story while in the back seat of a car	Single young boy's voice	

⁴ These source sequences will be made available free of charge for research purposes on the Consumer Digital Video Library (CDVL). The CDVL will be available at www.cdvl.org. This web site is scheduled to be deployed by the third quarter of 2009

Catjoke	Man tells a joke	Single man's voice	
Cchart2	Man describes the use of a color chart	Single man's voice	
Rfdev2	Man describes using a piece of technical equipment	Single man's voice	
Smity1	Man describes using a piece of technical equipment	Single man's voice	
Spectrum1	Man describes a radio spectrum chart	Single man's voice	
Vtclnw	Woman reads news copy	Single woman's voice	

The processed sequences for the study were produced using high quality (i.e. no additional impairments introduced) audio and CIF resolution video signals, with varying amounts of audio

offset introduced to produce each multimedia sample. The audio offset was varied between +405 and -405 milliseconds (i.e., up to approximately half a second before or after the video), at intervals of 45ms. Figure 2 shows the test sequence creation process. The audio and video were saved into separate files. Each audio was shifted in time using the 19 fixed offsets described above. Each video sequence was de-interlaced and format converted from the original resolution to CIF (352 pixels by 288 lines, 30fps). The video was then combined with each of the 19 different delayed or advanced audios. Finally, the first one second and last one second was discarded from each of the processed audio-video sequences. This resulted in 19 versions of each of the 9 source audio-video sequences, for a total of 171 test sequences, each 12 seconds in duration. The audio and video were kept in uncompressed formats throughout this process.

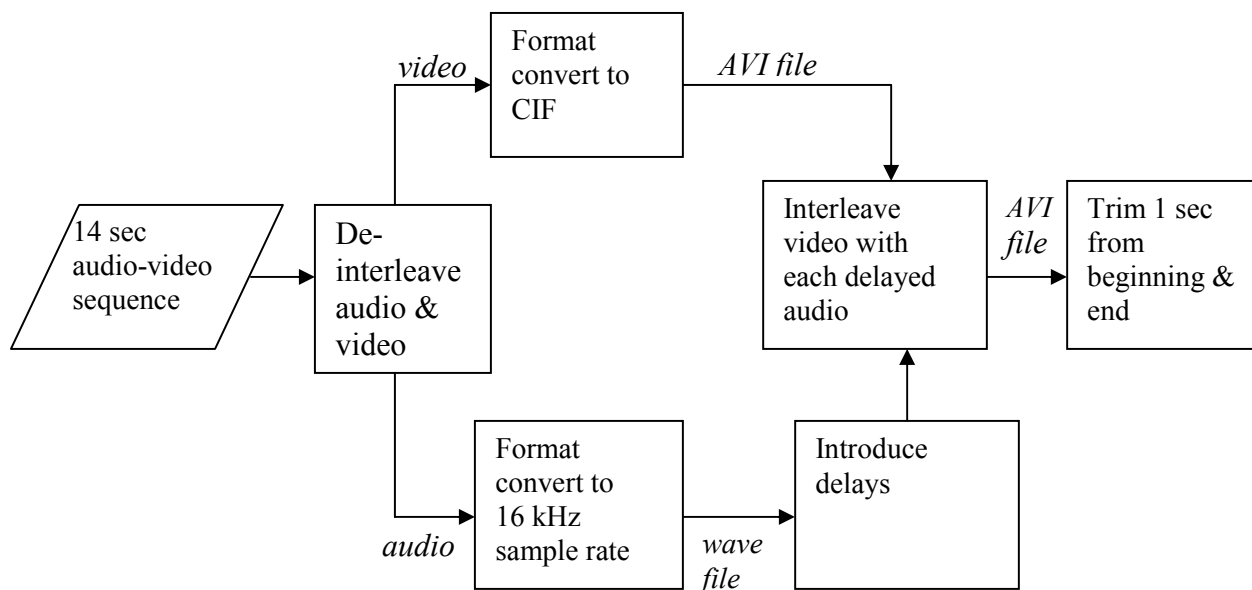


Figure 2. The sequence of creating testing samples

2.2 Subjective Test Method

The subjective test was performed using the single stimulus Absolute Category Rating (ACR) methodology as described in [6]. Test subjects rated each processed sequence on a scale of: excellent, good, fair, poor, and bad. Subjective ratings were gathered for 24 naïve subjects. Subjects rated all 171 test sequences. Each subject viewed a different randomization of the test sequences. They were allowed to take the test at their own pace and given a short break halfway through. For all subjects, the test took less than 45 minutes.

The viewing-listening environment was a sound-isolated chamber, and the subjects were allowed to view the sequences at a comfortable distance of their choosing. This was nominally 6-8 picture heights and was deemed typical of a desktop personal computer (PC) environment. The video was shown on an LCD monitor, in CIF format. The audio was played on two speakers placed on either side and behind the LCD monitor. The instructions read to the viewer are given in Appendix A, and the software used to administer the test is described in Appendix B.

3 DATA ANALYSIS

To ensure accurate control of the differential delay under study, the audio and video streams were tested to measure any differential delay between them introduced by the playback system. A constant video delay introduced by the test setup of 50 ms was identified. Further details on the analysis of this delay may be found in Appendix C.

Figure 3 contains a box plot that shows the data distribution for all audio-video sequences, plotted as a function of audio offset. This plot accounts for the 50 ms video delay described above. Thus, the range of audio offsets changes from -405 ms to +405 ms to -455 ms to +355 ms when taking into account the system delay. The bottom and top of each box indicate the 25th and 75th percentile, respectively. The bar in the middle of the box identifies the mean opinion score (MOS) for that audio offset, averaged across all scenes. The range spanned by the minimum and maximum MOS is drawn as a bar extending below and above the box, respectively.

An audio offset less than zero indicates that the audio leads the video. An audio offset greater than zero means that the audio is delayed in relation to the video. An audio offset of zero means that the audio and video were synchronized.

The MOS measurements were obtained by averaging the scores of all test scenes with the same offset. An interesting result of this test is that positive audio offsets (audio delayed in relation to the video) were rated to be of higher quality than the same negative audio offset, and MOS scores degraded more rapidly for negative audio offsets than for positive audio offsets. These data trends follow because of the difference between the speed of light and the speed of sound. Since light (i.e., video information) travels at a much higher rate than sound (i.e., audio information) people are accustomed to the situation where video information is received before audio information. Thus, a positive audio offset is perceived as less annoying and less out of the ordinary than a negative audio offset.

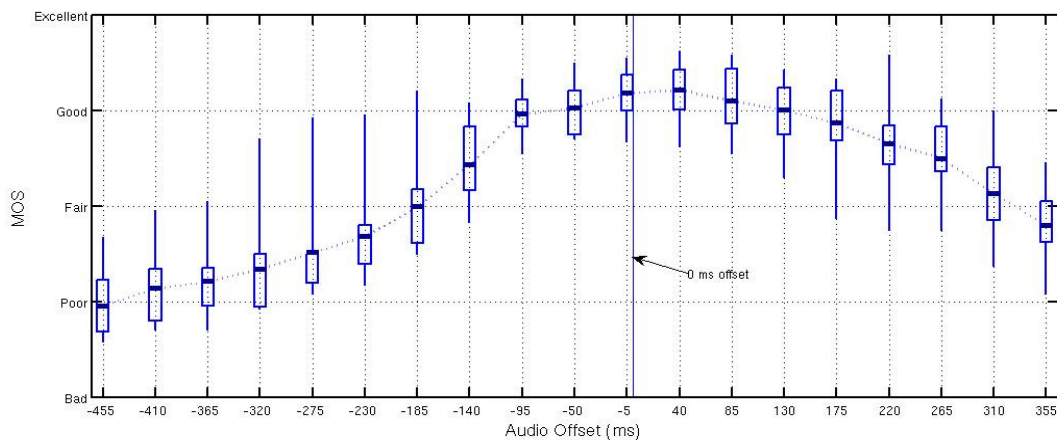


Figure 3. MOS as a function of audio offset

4 FUTURE WORK

It would be useful to perform the same test as described in this report with different audio-visual footage to examine whether the trends discussed in Section 3 are manifest with a different set of test scenes and, if so, what aspects of video might cause this observation; one example would be the effect of the relative size of the speakers' lips on the screen has on the perception of the differential delay.

The next step in the series of multimedia quality studies is in the area of HDTV. The first HDTV study will focus on the effects of the perceived audio and video quality on the combined perceived multimedia quality, with no differential delay.

5 REFERENCES

- [1] K. Brunnström, D. Hands, F. Speranza, and A. Webster, “VQEG validation and ITU standardization of objective perceptual video quality metrics,” *IEEE Signal Processing Magazine* [97], May 2009.
- [2] ITU-T Recommendation J.144, “Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference,” Geneva, 2003 (available at www.itu.org).
- [3] ITU-T Recommendation J.148, “Requirements for an objective perceptual multimedia quality model,” Geneva, 2003 (available at www.itu.org).
- [4] ITU-T Recommendation P.862, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs” Geneva, 2001.
- [5] ITU-R Recommendation BT.1359-1, “Relative timing of sound and vision for broadcasting,” Geneva, 1998.
- [6] ITU-T Recommendation P.910, “Subjective Video Quality Assessment Methods for Multimedia Applications,” Geneva, 2007.
- [7] ITU-R Recommendation BT.500, “Methodology for subjective assessment of the quality of television pictures,” Recommendations of the ITU, Radiocommunication Sector.

APPENDIX A: VIEWER INSTRUCTIONS

Thank you for coming in to participate in our study. The purpose of this study is to gather individual perceptions of the quality of several short [video / audio / multimedia] files. This will help us to evaluate various transmission systems for those files.

In this experiment you will be presented with a series of short [video / audio / multimedia] clips. Each time a clip is played, you will be asked to judge the quality of the clip. A ratings screen will appear on the screen and you should use the mouse to select the rating that best describes your opinion of the clip. After you have clicked on one of the options, click on the “Rate” button to automatically record your response to the hard drive.

Observe and listen carefully to the entire clip before making your judgment. Keep in mind that you are rating the [visual / audio / visual and audio] quality of the clip rather than the content of the clip. If the subject of the clip is pretty or boring or annoying, for example, please do not consider that when evaluating the physical quality of the clip. Simply ask yourself what you would think about the quality of clip if you [saw / heard] this clip on a [small television or computer or music player].

And don't worry about somehow giving the wrong answer; there is no right or wrong answer. Everyone's opinion will be slightly different. We simply want to record your opinion.

We will start with a few practice clips while I am standing here. After that, the experiment will be computer controlled and will be presented in two blocks of about 20 minutes each.

After the first block is finished, the computer will tell you that the section is finished. You should stand up and push open the door and come out of the chamber. By the way, the door will never be latched or locked. The door is held shut with magnets, much like modern refrigerators [demonstrate the pressure needed to push open the door]. If you have claustrophobia or need to take an unscheduled break, feel free to open the door and step outside for a moment.

During the break between sessions, there will be some light refreshments for you. When you are ready, we will begin the second session.

Do you have any questions before we begin?

APPENDIX B: TEST SOFTWARE DESCRIPTION

The multimedia testing software uses the same Java™ graphical user interface (GUI) for both the administrator and the subjects. The interface allows the administrator to customize the test. The administrator has control over options such as the desired video output drivers (DirectX or OpenGL), the ability to run the test on one or two monitors, the quality scale (five or nine level scales are available), and whether the subject is rating audio files, video files, or both. These settings can be saved so that later tests can be run with those identical settings. Figure B-1 shows a screen shot of the interface where the administrator specifies these variables.

The desired video and/or audio clips are loaded through the GUI by the administrator. Clips can be loaded into either the practice space (allowing users to get a feel for the testing procedure without the results being counted) or into the testing space (in which the user's ratings are registered and saved to a numbered file on the hard drive). The screen used to choose the audio-video files is shown in Figure B-2.

Once the test environment is created, pressing "Start Test" in Figure B-1 starts the subjective test interface. An introductory screen is presented while the viewer is seated. When the subject is ready to begin, an on-screen button is pressed to play the practice clips. The video and audio files are played in a random order using the freeware player MPlayer™, using command line calls from within the Java GUI. A different player can be used, provided that it has a command line interface and suitable GUI. After viewing and/or listening to the sequence, the subject chooses a rating based on what they saw and/or heard (see Figure B-3). After the practice session, the software pauses to allow questions to be asked. Then, the subject is presented with the audio-video sequences from the experiment.

The subject's opinion scores are saved to a file and associated with that subject's identification number. Subject identification numbers were not associated with subjects' names due to privacy concerns.

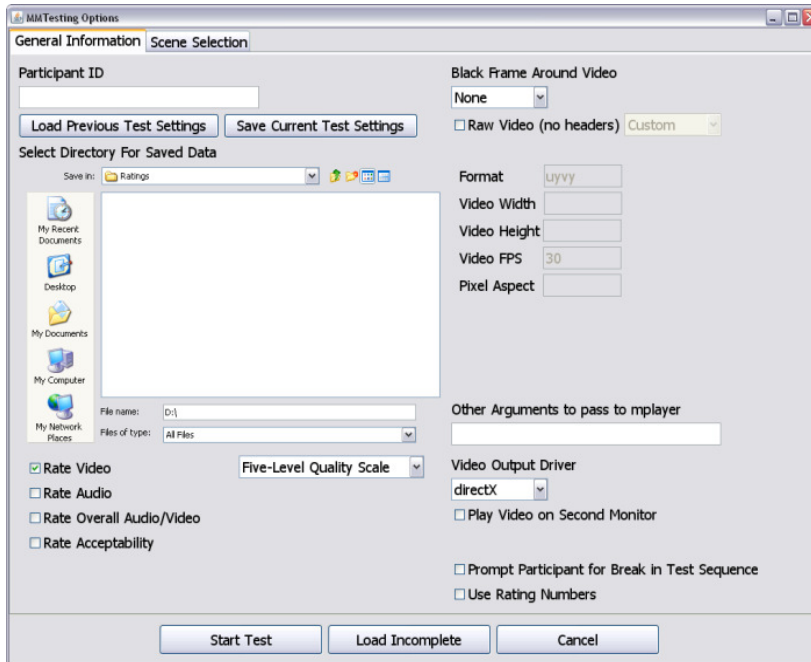


Figure B-1. Subjective test control program interface used to specify type of experiment.

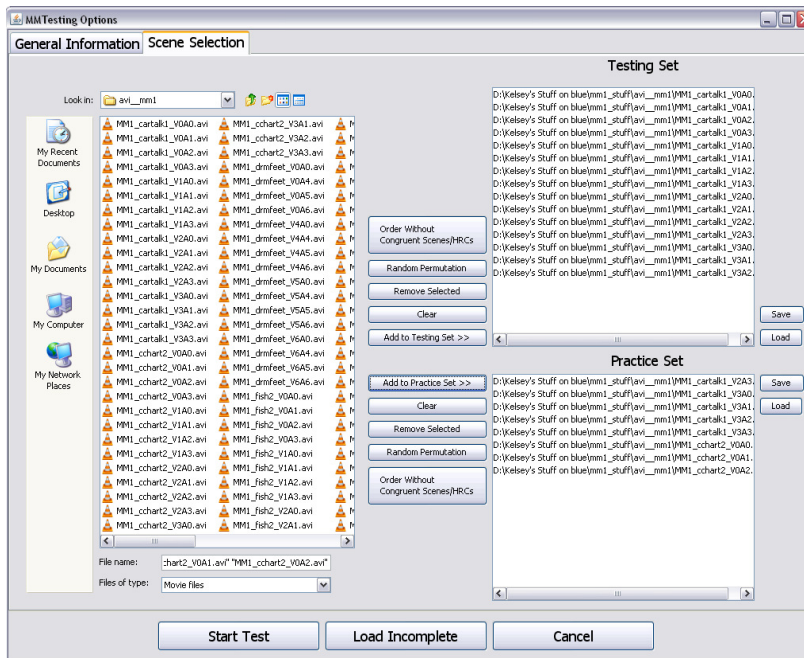


Figure B-2. Subjective test control program interface used to specify test sequences and practice session.

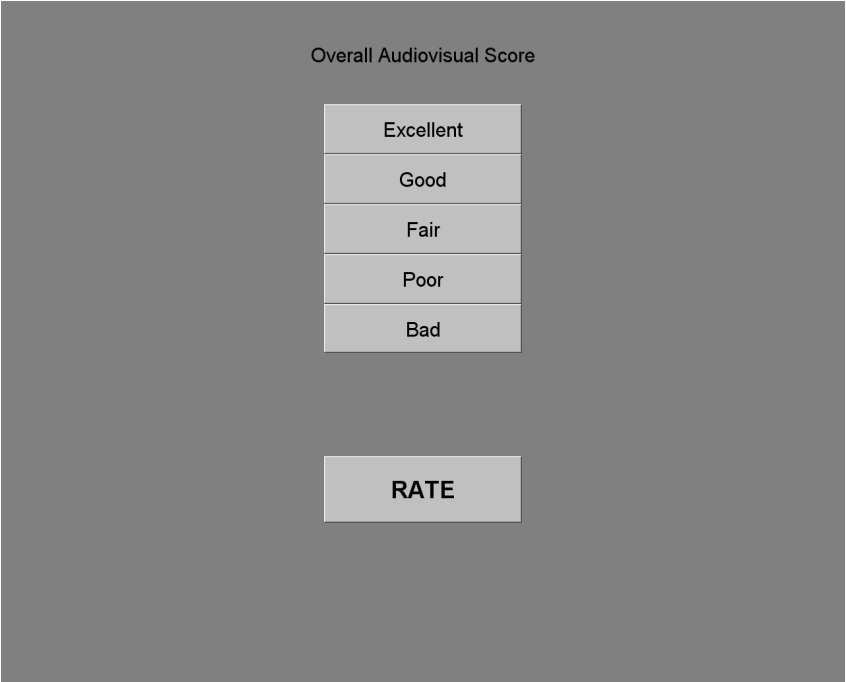


Figure B-3. The rating screen showing the MOS scale.

APPENDIX C. AUDIO/VIDEO CALIBRATION

To ensure accurate control of the differential delay under study, the audio and video streams were tested to measure any differential delay between them introduced by the playback system.

The validation test was designed to mimic the actual subjective test as closely as possible. Two computers were used, one to play the video and the other to capture the results. The results were captured using a specially designed AVI video sequence and a photoelectric sensor. The video signal was uncompressed UYVY with a resolution of 352 by 288 at 30 frames per second stored in an AVI format. Every frame in the video was a simple black frame, with the exception of a single white frame every two seconds. This corresponded to the audio track that produced a 440 Hz tone for exactly 33.3 milliseconds ($1/30$ of a second or one frame) every two seconds. Figure C-1 demonstrates the test setup.

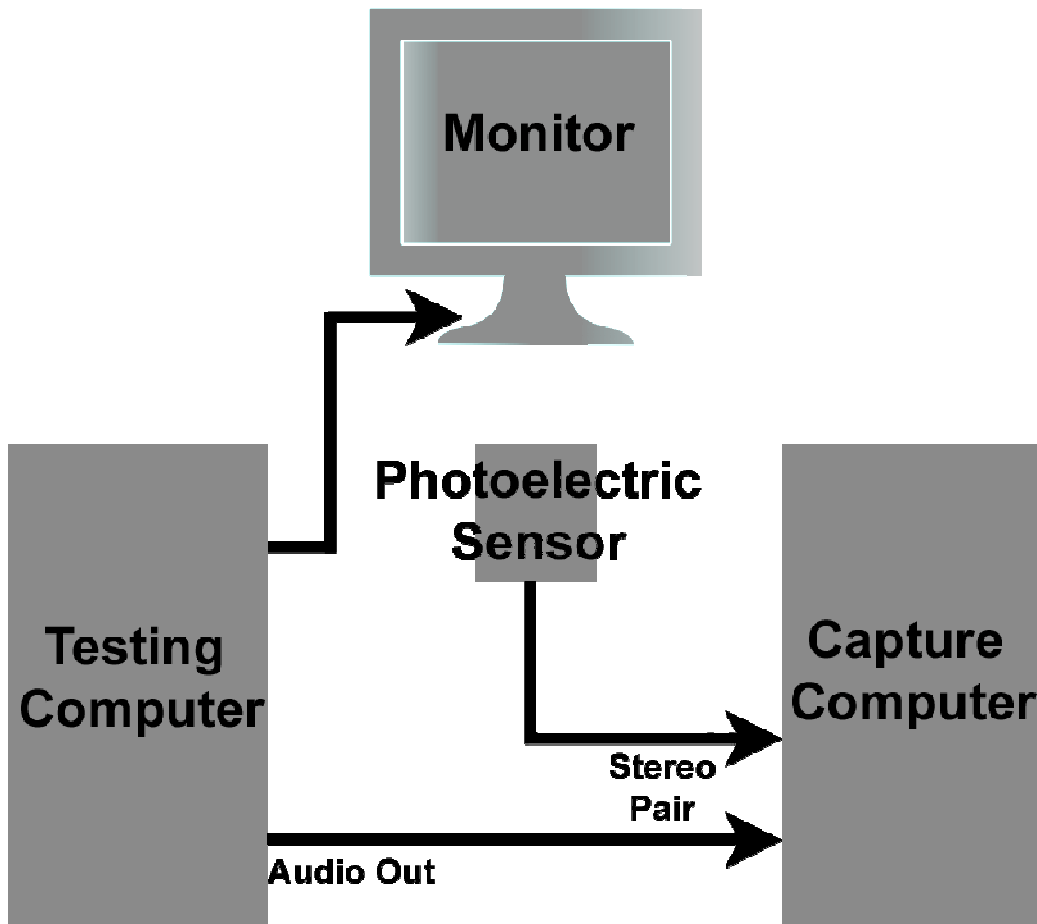


Figure C-1. The equipment setup for the data validation test.

The computers were set up so that the testing computer played the video and displayed it to a monitor in a darkened room. As the video played, the white frames were displayed and detected by the photoelectric sensor. The sensor and the audio output from the test computer formed a stereo signal that was fed to the capture computer's audio input. In this way, the video could be interpreted as one channel of a stereo audio signal and directly compared to the other channel, which was the audio from the video. Testing in this manner allowed a direct correlation to be drawn between the audio and video synchronization.

The validation test was run on two separate test computers. The test was run using both MPlayer and VLC Media Player™. MPlayer™ is the media player that is currently used by the subjective testing software, and was configured in an identical way to a typical subjective test. VLC was used as a control. There was no significant difference between the media players; each had similar differential delays.

The validation test was run for two minutes. The results indicate that there was a 50 ms delay between the video signal and the audio signal, where the video lagged the audio. The cause of this delay is likely due to the video processing computer configuration.

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION NO. TM-10-464		2. Government Accession No.	3. Recipient's Accession No.
4. TITLE AND SUBTITLE Multimedia Synchronization Study		5. Publication Date October 2009	
		6. Performing Organization Code	
7. AUTHOR(S) Carolyn Ford, Mark A. McFarland, William Ingram, Scott Hanes, Margaret Pinson, Arthur Webster, Kelsey Anderson		9. Project/Task/Work Unit No. 3139	
8. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Telecommunication Sciences National Telecommunications & Information Administration U.S. Department of Commerce 325 Broadway Boulder, CO 80305		10. Contract/Grant No.	
		12. Type of Report and Period Covered Tech Memorandum, FY09	
11. Sponsoring Organization Name and Address National Telecommunications & Information Administration Herbert C. Hoover Building 14 th & Constitution Ave., NW Washington, DC 20230			
14. SUPPLEMENTARY NOTES			
15. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.) ITS is conducting a series of studies to quantify the effects of the separate audio and video transmission qualities, and the differential delay in their synchronization, to the perceived aesthetic quality of a multimedia signal. The experiment described in this report was designed to specifically study the effects of the differential delay.			
16. Key Words (Alphabetical order, separated by semicolons) audio delay; audio offset; audio-video synchronization; differential delay; multimedia quality; subjective testing			
17. AVAILABILITY STATEMENT <input type="checkbox"/> UNLIMITED.		18. Security Class. (This report) Unclassified	20. Number of pages 14
		19. Security Class. (This page) Unclassified	21. Price:

NTIA FORMAL PUBLICATION SERIES

NTIA MONOGRAPH (MG)

A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA SPECIAL PUBLICATION (SP)

Conference proceedings, bibliographies, selected speeches, course and instructional materials, directories, and major studies mandated by Congress.

NTIA REPORT (TR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities. Subsets of this series include:

NTIA RESTRICTED REPORT (RR)

Contributions that are limited in distribution because of national security classification or Departmental constraints.

NTIA CONTRACTOR REPORT (CR)

Information generated under an NTIA contract or grant, written by the contractor, and considered an important contribution to existing knowledge.

JOINT NTIA/OTHER-AGENCY REPORT (JR)

This report receives both local NTIA and other agency review. Both agencies' logos and report series numbering appear on the cover.

NTIA SOFTWARE & DATA PRODUCTS (SD)

Software such as programs, test data, and sound/video files. This series can be used to transfer technology to U.S. industry.

NTIA HANDBOOK (HB)

Information pertaining to technical procedures, reference and data guides, and formal user's manuals that are expected to be pertinent for a long time.

NTIA TECHNICAL MEMORANDUM (TM)

Technical information typically of less breadth than an NTIA Report. The series includes data, preliminary project results, and information for a specific, limited audience.

For information about NTIA publications, contact the NTIA/ITS Technical Publications Office at 325 Broadway, Boulder, CO, 80305 Tel. (303) 497-3572 or e-mail info@its.blrdoc.gov.

This report is for sale by the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161, Tel. (800) 553-6847.

