Document Title:	Research on Videoconferencing for Pretrial Release Hearings, Version 1.0
Author(s):	Johns Hopkins University
Document Number:	252945
Date Received:	May 2019
Award Number:	2013-MU-CX-K111

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

11100 Johns Hopkins Road • Laurel, Maryland 20723-6099

AOS-18-0873

NIJ RT&E Center Project 16–2

December 2018

RESEARCH ON VIDEOCONFERENCING FOR PRETRIAL RELEASE HEARINGS

Version 1.0

Prepared for:

NIJ | National Institute of Justice

STRENGTHEN SCIENCE. ADVANCE JUSTICE.

Prepared by:

The National Criminal Justice Technology Research, Test, and Evaluation Center
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Rd.
Laurel, MD 20723-6099

Task No.:	FGSGJ
Contract No.:	2013-MU-CX-K111/115912

This project was supported by Award No. 2013-MU-CX-K111, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

The research described in this report was sponsored by the National Institute of Justice, prepared and conducted by The Johns Hopkins University Applied Physics Laboratory (JHU/APL).

Published by The Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland

For more than 70 years, JHU/APL has provided critical contributions to critical challenges with systems engineering and integration, technology research and development, and analysis. As the Nation's largest University-Affiliated Research Center, JHU/APL undertakes vital national security and scientific challenges in a way that is free from conflicts of interest or competition with commercial industry.

www.jhuapl.edu

Task No.: FGSGJ
Contract No.: 2013-MU-CX-K111/115912
Tracking No.: AOS-18-0873

This project was supported by Award No. 2013-MU-CX-K111, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect those of the Department of Justice.

ACKNOWLEDGMENTS

The preparation of this report required the dedication of many people. The study team is grateful to all the individuals whose expertise and prior work provided the foundation for this study. These people include the researchers of the *Research on Videoconferencing at Post-Arraignment Release Hearings: Phase 1 Final Report*, under NIJ award number GS-23F-8182H and published in June 2015, available at https://www.ncjrs.gov/pdffiles1/nij/grants/248902.pdf. The co-authors were Robin Davis, Ph.D.; Billie Jo Matelevich-Hoang, J.D.; Alexandra Barton, B.A.; Sara Debus-Sherrill, M.A.; and Emily Niedzwiecki, M.S. In addition, the team would like to extend gratitude to all the reviewers whose critical and constructive feedback greatly improved the quality of the report.

Finally, the study team would like to thank National Institute of Justice Program Manager, Martin Novak, for his support throughout this study and the preparation of this report.

CONTENTS

FIGURES

TABLES

EXECUTIVE SUMMARY

Numerous studies related to the effectiveness of video teleconferencing (VTC) in courts have identified differences in the outcomes of hearings conducted using VTC compared to traditional courtroom hearings. Some studies suggest that the use of VTC technology may unfairly bias proceedings. While it is exceedingly difficult to isolate the cause of these differences in a complex system involving hardware, software, information transmission, and human perception, eliminating quantifiable elements from consideration would help to identify causal elements.

This study examined the effects of file degradation typical of network issues specifically to evaluate the usefulness of objective and subjective methods of measuring video quality and the impact of packet loss and jitter (latency) on the perceived effectiveness of VTC hearings.

The overall objective of the National Criminal Justice Technology Research, Test, and Evaluation Center research was to conduct experimental quantitative and qualitative research to determine whether quantitative metrics for video quality could be found that correlate highly with human subjects' perceived video quality.

The study team identified reference videos that would test conditions that might affect the VTC system's ability to digitally capture and display scenes reasonably found in a courtroom, which in turn might affect viewer's perception of demeanor, including variation in contrast between subject and the background, initial white balance, angle of the lighting on the subject, distance or angle between camera and subject, and reflective surfaces. A total of 138 clips representing various levels of introduced jitter and packet loss were created.

Three quantitative metrics were selected to objectively evaluate the quality of the sample videos: perceptual evaluation of video quality (PEVQ) and Structural SIMilarity (SSIM), which are based on human perception models; and peak signal-to-noise ratio (PSNR), which is not human perception based. Then using a five-point Likert-type scale, human subjects rated their perceptions of the same video clips for subjective video quality corresponding to bad (1), poor (2), fair (3), good (4), excellent (5); and their ability to interpret facial expressions, using a scale ranging from 'cannot at all interpret' (1) through 'can very easily interpret' (5) the facial expression.

The study found that participants believed they were able to adequately discern facial expressions of the subjects in the video despite noticeable levels of jitter and packet loss distortions. Thus, there was utility in videos with some levels of apparent noise. Therefore, subjective rating of utility should be explored further to determine an acceptability threshold for automated video quality assessment tools.

While more study is needed, objective measures appear to be more conservative than human participants in the scoring of video. As a result, if the objective measure determines that video is bad or poor, there is a high likelihood the video is not usable. If the objective measure determines the video is fair, good, or excellent, there is a high likelihood the video is usable.

Based on the combined objective and subjective measures analysis, all three objective tools appear to be acceptable alternatives to subjective measures. The rank order of goodness-of-fit for the objective tools are:

1. PEVQ
2. PSNR
3. SSIM

"If video quality is poor, this could influence or affect the hearing."[1]

Research on Videoconferencing at Post-Arraignment Release Hearings:
Phase I Final Report, May 2015

1. INTRODUCTION

Despite the benefits of using video teleconferencing (VTC) technology in courts to increase efficiency and access to the justice system while reducing pretrial detention time and transportation costs,[2] some studies suggest that the use of VTC technology may unfairly bias proceedings. Numerous studies related to the effectiveness of VTC in courts identified differences in the outcomes of hearings conducted using VTC when compared to traditional courtroom hearings.[3]

This study focused on evaluating the usefulness of objective and subjective methods of measuring video quality and the impact of network issues on the perceived effectiveness of VTC hearings.

1.1 BACKGROUND

In September 2013, the Johns Hopkins University Applied Physics Laboratory (JHU/APL) was selected by the Department of Justice, National Institute of Justice (NIJ) to establish the National Criminal Justice Technology Research, Test, and Evaluation Center (RT&E Center) within the National Law Enforcement and Corrections Technology Center System. The purpose of the RT&E Center is to provide in-depth technical reports and support for NIJ's non-forensic research and development efforts.

A previous study entitled, "Research on Videoconferencing at Post Arraignment Release Hearings: Phase 1 Final Report" (referred to in this report as Phase 1 Final Report) was conducted under NIJ Award GS-23F-8182H and published in 2015. The RT&E Center was invited to conduct a follow up project under Award 2013-MU-CX-K111. This report documents the RT&E Center's study entitled, "Research on Videoconferencing Technology at Pretrial Hearings."

The NIJ videoconferencing project Phase I Final Report[4], published in 2015, noted that potential benefits to the use of VTC include decreased staff/personnel travel time to and from detention and court settings; decreased costs of transporting inmates from detention to court settings;

[1] Davis et al. "Research on Videoconferencing."
[2] Davis et al. "Research on Videoconferencing."
[3] Diamond et al. "Efficiency and Cost."
[4] Davis et al. "Research on Videoconferencing" 15.

increased security of inmates and detention and court staff; decreased medical costs due to reduced exposure to other inmates in close confines of transport; reduced offsite meal costs; and increased overall efficiency of procedural hearings.

And yet the report also noted that the participants who contributed to the study expressed conflicting opinions in favor of or against the use of videoconferencing technology based on their experiences and expertise. An internet search for "video teleconference hearing" results in scores of websites suggesting that people should refuse such hearings because:

> *A VTC hearing could obscure non-verbal communication which in turn "could negatively influence your credibility and jeopardize the outcome of your claim."*[5]
>
> *"Many argue, that without the ability of the Judge to physically observe and assess the claimant, the probability increases that the decision will be unfavorable."* [6]
>
> *"Credibility is an important aspect to most disability cases, and it is easier for a Judge to assess credibility in-person rather than through a monitor."*[7]
>
> *"most claimants would do better at an in-person hearing than a video hearing" because "exceptional video quality is required to capture subtle facial expressions" and other non-verbal communication.* [8]
>
> *"Our win rate is approximately 5% higher with live judges than with video judges."*[9]

The claims above were posted by legal firms who did not include supporting data, however; some scientific studies suggest that the use of VTC technology may unfairly bias the perceived credibility of the testimony when compared to traditional courtroom hearings. There is a long case history of judges relying on observation of demeanor when determining credibility.[10] Demeanor may include the subjects' appearance, behavior, and tone of voice. Wellborn stated, "According to the empirical evidence, ordinary people cannot make effective use of demeanor in deciding whether to believe a witness. On the contrary, there is some evidence that the observation of demeanor diminishes rather than enhances the accuracy of credibility judgments." Nonetheless, a survey of trial judges found that credibility was most often based upon "evasiveness, defensiveness, and rationalization" indicated by changes in the witness's behavior.[11] Therefore, if the use of VTC prevents the judge from observing these indicators, there would be a direct impact on the ability to evaluate credibility.

In a study of the use of VTC for bail hearings in Cook County, Illinois (Chicago), Diamond *et al* reported assigned bail was 51 percent higher on average in remote hearings using VTC than in

[5] Truehelp (web page).
[6] Syfrett, Dykes, & Furr (web page).
[7] Smith (web page).
[8] Petow (web page).
[9] Quikaid, (web page).
[10] Timony, "Demeanor Credibility."
[11] Timony, "Demeanor Credibility."

live hearings with defendants present.[12] In another study, investigators reported higher levels of removal in remote deportation hearings, raising questions about the fairness of the technology to defendants in immigration hearings[13]. And in a much-cited study, Goodman, et al.[14] concluded that jurors believed children testifying via VTC to be less creditable than children testifying live in court, leading to concerns that defendants interacting via video are seen as less credible than defendants who are present in the courtroom. All of these studies suggest a negative bias toward participants in a remote hearing.

Such effects could be due to inherent differences in the way humans perceive or interpret digital representations of the input they would otherwise receive directly through their own senses. The *Phase 1 Final Report* found that aspects related to how well participants could see or hear others in the court proceeding, including "who is in view, the nuanced interactions between individuals, and facial expressions ... impact the decorum of the court and whether the hearing is experienced similarly to an in-person hearing."[15] The report further indicated that that screen resolution may cause "nonverbal cues or body language to be missed or misinterpreted by courthouse parties." Similarly, sound quality and control are important to ensure that parties can hear softer sounds "such as the defendant muttering that he/she does not understand something or wants to speak."[16] Raising concerns that aspects of the video itself could impact the ability to determine demeanor and thus impact perception of credibility.

The potential impact of video quality on the outcome of a VTC hearing is indicated by another study, in which a judge stated that the ability to observe a participant's demeanor and emotions was a deciding factor when determining whether VTC was an acceptable alternative to an in-person hearing.[17] And a case brief included the judge's conclusion supporting use of a VTC hearing, stating that the VTC video quality was "flawless" and that "any hesitation, discomfort, arrogance, or defiance would have been easily discerned."[18] And yet, Lassiter found that the camera angle used to video record a confession impacted the viewers' opinion of whether the confession was voluntary, and thus whether they believed the defendant was guilty.[19] Suggesting that even with a clear image, other factors may still contribute toward a bias against VTC testimony.

Acceptance of VTC hearings requires research-based guidance about which situations are appropriate for the use of VTC hearings and which are counter-indicated. Therefore, fully understanding quantifiable elements in a complex system involving hardware, software, information transmission, and human perception is a critical step to investigating the possible physiological, psychological, and sociological effects on the outcomes of VTC hearings.

[12] Diamond et al. "Efficiency and Cost."

[13] Haas, "Videoconferencing in Immigration Proceedings."

[14] Goodman et al., "Face-to-face Confrontation."

[15] Davis et al. "Research on Videoconferencing," 24.

[16] Davis et al. "Research on Videoconferencing," 24.

[17] Davis et al. "Research on Videoconferencing," 15.

[18] U.S. Court of Appeals, Case No. 15–1349, Document No. 1613347, filed: 05/16/2016. 70.

[19] Lassiter et al. "The Potential for Bias in Videotaped Confessions," 1838–1851.

1.2 OBJECTIVES

The RT&E Center study was designed to build upon the foundation established in the Phase 1 Final Report and further guide the use of VTC in pretrial hearings. Specifically, the study objective was to evaluate the effects of network degradation on video quality through objective and subjective testing in an effort to:

- Identify and assess methods, metrics, and tools for measuring video quality; and
- Determine if there is an objective measurement of video quality which correlates with subjective measures of ability to perceive facial expressions, which is used to model whether the video quality of a remote hearing conducted using VTC is 'sufficient' for a VTC hearing.

1.3 STUDY TEAM

The study team members contributed a range of skills to the project.

Team Member	Role / Expertise
Eliud Bonilla	Digital Multimedia Forensics
Lauren Brush	Team Lead, Human Factors Analyst, PMP
Jay Chang	Electrical Engineer
John Cristion	Team Lead, Signal Processing Engineer
Aaron David	Computer Engineer, Network Design
Steven Ferraro	A/V Systems Design Engineer
Stuart Fogel	Systems Engineer, Member MD State Bar, ABA
Mark Gabriele	Project Manager
Jesse Gruter	Legal Policy Advisor, Member TX State Bar, ABA
Kelly O'Brien	Subjective Test Lead
Robert Pattay	Electrical Engineer, Video Quality Assessment
Matthew Spisso	Software Engineer, Test & Evaluation
Daniel Syed	Objective Test Lead, Systems Engineer
David Vespoint	Video Engineer
Richard (DJ) Waddell	Program Manager

2. METHODOLOGY

The overall procedure consisted of using a simulated teleconferencing system (Figure 1) to mimic network induced distortion likely to impact resulting video quality. The video quality of test videos was measured using both objective and subjective methods, and then correlation between the objective and subjective results was investigated.

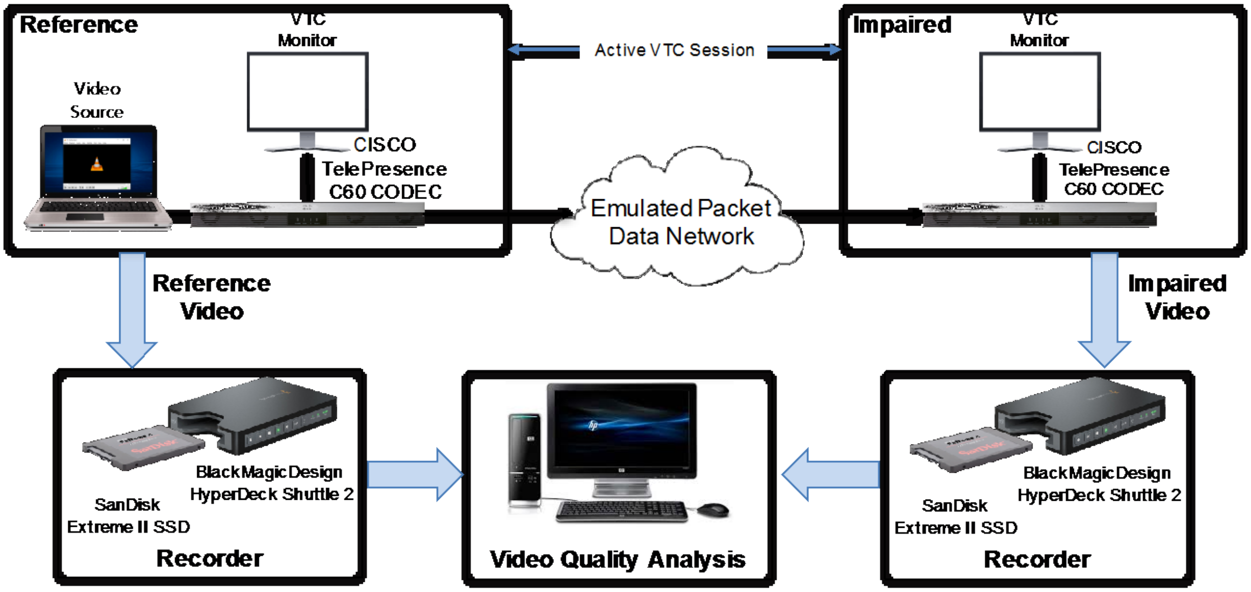


Figure 1: VTC Testing Environment

The following section gives a summary of the methods used in this study. Details of each step leading to the correlation between objective and subjective results are documented in separate appendices as follows:

1. Measuring Audio and Video Quality – see Appendix A
2. Video Selection and Preparation – see Appendix B
3. Video Quality – Objective Assessment – see Appendix C
4. Video Quality – Subjective Assessment – see Appendix D
5. Subjective Assessment Raw Data – see Appendix E

2.1 TEST BED DESIGN

The RT&E Center first investigated existing work on VTC efficacy in pretrial hearings and measuring VTC quality. Much of the discovery effort focused on characterizing VTC system components, understanding of the situations and conditions that may limit their effectiveness, and identifying VTC quality metrics and measurement tools.

The discovery process guided the subsequent design and execution of the test and measurement steps. For example, based upon the discovery phase, the RT&E Center's Test Plan assumed that VTC system hardware, software and codecs were unlikely to have a negative impact on VTC hearings. Furthermore, when setup and operated according to established best practices, the VTC environment is not expected to have a detrimental effect upon the court proceedings.

JHU/APL's Advanced Networking Technologies Lab Hardware in the Loop Test Bed (ANT-HIL) consists of both physical and virtualized networking hardware and functionality. It is a Linux-based environment hosted on a Dell PowerEdge R720 server running VMware ESXi and allows the creation of various emulated packet data networks with customizable topologies and characteristics. Additionally, the ANT-HIL server is connected to a Juniper EX4200 switch to allow connections between the virtualized networks, physical hardware, and other networks. Cisco Tandberg C60 VTC endpoints[20] are connected to the Juniper EX4200 switch through a Cisco Catalysts 3750G switch located in the ANT-HIL lab, and packets sent between the two VTC endpoints are routed through the emulated network. Figure 2 represents the network topology used for the emulated environment.

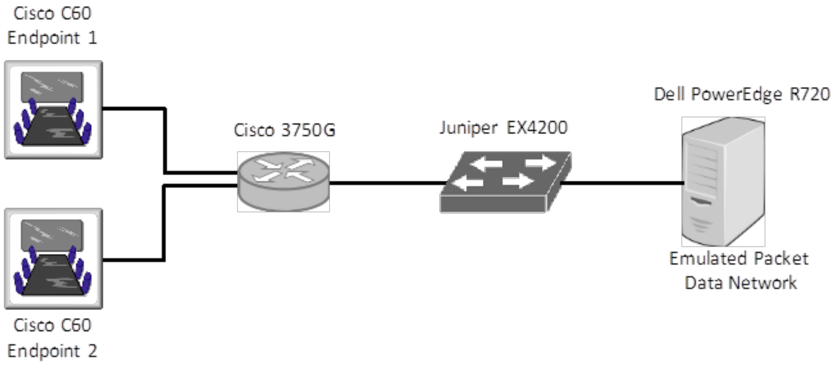


Figure 2: Emulation Network Topology

The emulated packet data network uses the network emulation (NetEM) kernel module to emulate wide area network (WAN) link characteristics. Network impairment characteristics, such as delay, packet loss, and other variables can be added to outgoing packets on any interface in the emulated network.[21] This allows hardware under test to be physically collocated yet still appear to be communicating over a WAN or other realistic network topology. Controlled testing in this environment provides insight into the effects of the network characteristics on end-to-end application performance.

2.1.1 Measuring Audio and Video Quality

There are two basic types of measurements made for both audio and video: objective and subjective. Objective measurements tend to be performed at the signal level. Differences in the input and output signals are measured, thus objective measurements measure the performance of the delivery system. In contrast, subjective measurements require subjective assessments from human observers. Subjective measurements are qualitative and tend to measure the quality of the product. It is generally accepted that the most accurate measurements of audio or video quality require evaluation from human subjects.[22]

Because perception and biases differ from one person to the next, deriving statistically valid conclusions can be challenging. International Telecommunications Union-ITU Telecommunication Standardization Sector (ITU-T) standards recommend that subjective evaluations utilize at least four observers, with 10–15 observers preferred. The presence of 10–15 trained observers assessing the quality of a videoconferencing stream in a courtroom is likely

[20] Cisco, "Jitter and Network Delay."
[21] SysTutorials, "tc-netem (8) – Linux Man Pages."
[22] Huynh-Thu et al. "Study of Rating Scales."

to be disruptive as well as prohibitively expensive. In addition, concerns about the safety and welfare of human participants involved in behavioral research and the need to maintain privacy and confidentiality of some proceedings further discourage the options of conducting routine subjective quality assessments of courtroom VTC systems.

Objective measurements of video or audio quality usually derive a quality measurement based upon characteristics of the input and output signals, although some tools incorporate models of human perception. Also, because they are automated, they do not require large numbers of trained observers to sit and watch video and assess its quality for hours on end. As a result, objective measurements are in general repeatable and they provide quantitative results that support comparison. Automated objective measurement techniques are thus less intrusive and less costly than subjective measurements in a courtroom setting.

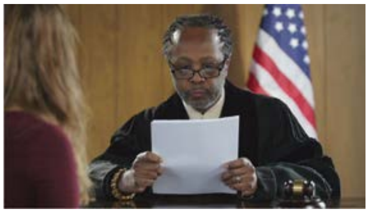
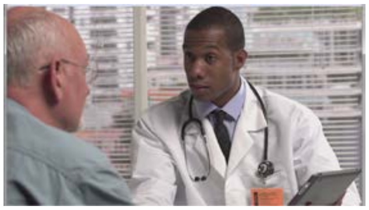
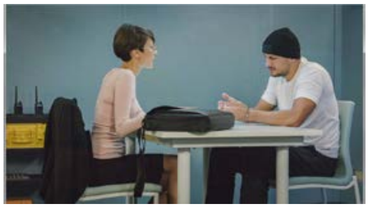
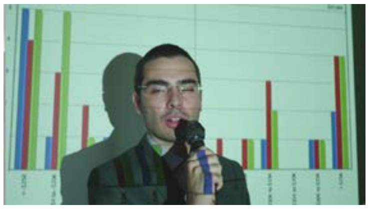
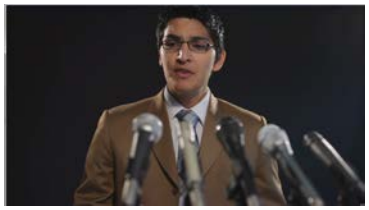
However, objective measures of audio video quality have their own limitations. First, the level of packet loss or delay in a signal may not have a predictable effect on a human observer. Two signals with equal levels of distortion can have significantly different value to human observers, depending on which packets are distorted and the distribution of distortion within an image. Thus, assigning meaning to an objective measure of quality to fit a large number of varying settings and use cases can be challenging. Second, the objective assessment of audio and video quality of the delivered signal through a simulated packet data network may not capture the effects of factors beyond the network. The audio and video captured in environments with insufficient lighting, poor acoustical pickup, or excessive background noise may impact the objective measurements and may be of limited value due to poor video quality caused by environmental factors and audio/video equipment limitations.

2.1.2 Video Selection and Preparation

Given the need to evaluate a variety of conditions, the option of obtaining actual courtroom video was eliminated due to difficulty locating multiple jurisdictions willing to share actual courtroom video, particularly video not filmed using best practices. The study team elected to purchase stock footage representative of courtroom situations, such as subjects seated or standing, addressing the camera, with limited movement, and no background movement. The study team selected video clips that would test conditions that could affect a viewer’s perception of demeanor and might reasonably be found in a courtroom.

The VTC equipment is less efficient than human eyes at capturing details in areas that are much brighter or much darker than the overall image. This can cause the facial features to be lost when subjects with dark skin are filmed against a bright background. Conversely, reflections appear as bright areas without detail, thus potentially obscuring the eyes of subjects wearing eyeglasses. Contrast was therefore identified as the primary test condition for study due to the known challenge that auto-contrast adjustment presents for VTC equipment and the potential impact on observing facial features necessary to interpret demeanor. Although best practices for room lighting and camera angle for VTC hearings exist, non-compliant videos were included to study whether a specific level of induced degradation caused an equivalent drop in quality when applied to videos filmed under both good and poor practices. The study was conducted using five reference videos, which captured the test conditions described in Table 1.

Table 1: Screenshots, Descriptions and Test Conditions Associated with Each Reference Video

Clip ID		Primary Test Condition	Secondary Test Condition(s)
A		Moderate contrast: dark skin tone / moderate background	Reflections Facial hair
F		High contrast: dark skin tone / light background	Oblique facial orientation Striated background
G		Moderate contrast: light skin tone / medium background	Increased distance between camera and subject
H		Low contrast: medium skin tone / medium background	Poor white balance Improper lighting angle
I		Moderate contrast: medium skin tone / dark background	Improper lighting angle Foreground objects

Several types of video distortion can occur when video is transmitted over a computer network. Jitter and packet loss were selected for this study because they each have a visible impact on video quality. Corruption of synchronization signals or electromagnetic interference during video transmission cause video jitter which is exhibited by randomly displaced horizontal lines in the video image frames. Jitter can be measured in milliseconds (ms) during which the data is corrupted. Packet loss occurs when one or more packets of data travelling across a computer network fail to reach their destination, causing areas of the image to appear to be missing. It is typically caused by network congestion but can also have other causes. Packet loss is often measured as a percentage of packets lost with respect to packets sent.

The test bed was designed to allow the capture of video files before and after transport over a simulated network. The simulated network enabled the insertion of controlled amounts of file

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

degradation to the recorded video clips. A 10-second section of each of the five reference videos (A, F, G, H, and I as shown in Table 1) was created in MP4 format as a baseline clip without any induced distortions. Eleven different levels of jitter and ten different levels of packet loss were chosen resulting in the following ranges:

- Jitter: ranged from 0 – 200 msec, at 20 msec intervals
- Packet loss: ranged from 0 – 50%, at 5% intervals

Also included in each series of video clips was a control video clip with no applied jitter or packet loss (distinct from the clips with a specified 0 distortion level). Because the test bed applies distortion randomly, the distortion process was performed twice on reference Clip A to allow comparison of two runs of the distortion process (series A and series Z) against a controlled starting video (reference Clip A). This resulted in six series of clips, each with 23 distinct versions, for a total of 138 total clips. These video clips were used for both objective and subjective video quality evaluation.

2.1.3 Video Quality – Objective Assessment

Based upon the Literature Review and prior experience at JHU/APL, the study team selected three objective tools: SSIM, PEVQ, and PSNR. Two of these tools, SSIM and PEVQ, attempt to measure video quality in a manner consistent with human perception. The third, PSNR, provides a simple, but readily repeatable and well understood metric, based on measurable network metrics, without a human visual model.

SSIM is a perceptual metric that quantifies image quality degradation caused by processing, such as data compression or losses in data transmission. It is a full reference metric that requires two images from the same image capture—a reference image and a processed image.

PEVQ is a full-reference perceptual measurement algorithm that performs pixel analysis of corresponding frames within two videos to generate an assessment of the perceptual quality of the output video. Degradations and artifacts resulting from coding of the video for network transmission are assessed using models of human visual perception. Results of these analyses are converted into a mean opinion score (MOS), which have been benchmarked against subjective assessments from human subject testing. PEVQ became part of the ITU-T Recommendation J.247[23] in 2008.

PSNR is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. The signal in this case is the original data, and the noise is the error introduced by compression. When comparing compression codecs, PSNR is an approximation to human perception of reconstruction quality.

More detailed descriptions of these tools and their use are included in Appendix A – Measuring Audio and Video Quality.

[23] ITU-T Rec. J.247.

2.1.4 Video Quality – Subjective Assessment

The same video clips were also shown to participants who were asked to subjectively rate the video quality. The subjective results were compared to the objective results in an effort to determine if any of the automated tools offer a reliable model for measuring video quality as it is perceived by humans.

Subjective testing employed human participants to rate their perception of overall video quality and also score their perceived ability to interpret facial expressions in the test videos as a measure of impact of the file degradation on interpretation of demeanor.

2.1.5 Correlation of Objective and Subjective Measurements

Finally, the results from the objective and subjective tests were analyzed for statistical correlation. The purpose of the analysis was to assess whether the method was repeatable for measuring the effect of VTC implementation on courtroom proceedings. If so, it could support future studies to evaluate the human factors associated with the effectiveness of VTC hearings as compared to traditional pretrial hearings, and thus, frame the protocols for appropriate use of VTC for hearings.

3. RESULTS

Summaries of the results and findings from the objective and subjective video quality assessments are included below, followed by results of the analysis of correlation between the two assessment methods.

3.1 VIDEO QUALITY – OBJECTIVE ASSESSMENT

When the initial objective video quality scores were determined for the Series A videos, each of the objective tools indicated that video quality was relatively steady for jitter rates of 0 through approximately 80 ms, whereupon video quality scores decreased. However, all tools showed an increase in video quality scores around 160 ms of jitter (Figure 3). To investigate this unexpected trend, a second series of degraded videos was generated from reference video A, called Series Z. Like Series A, Series Z also exhibited steady scores for several levels of jitter before the scores dropped, followed by a rise in scores indicating improved video quality, then a second drop in quality. Ultimately, jitter scores for each video series showed this pattern of steady initial quality scores, followed by a drop, then a temporary increase in quality scores.

The team concluded that this trend indicates that the VTC system compensates for reduced bandwidth, leading to an apparent increase in video quality scores. Video quality scores for packet loss scores were generally much steadier, making the pattern difficult to distinguish.

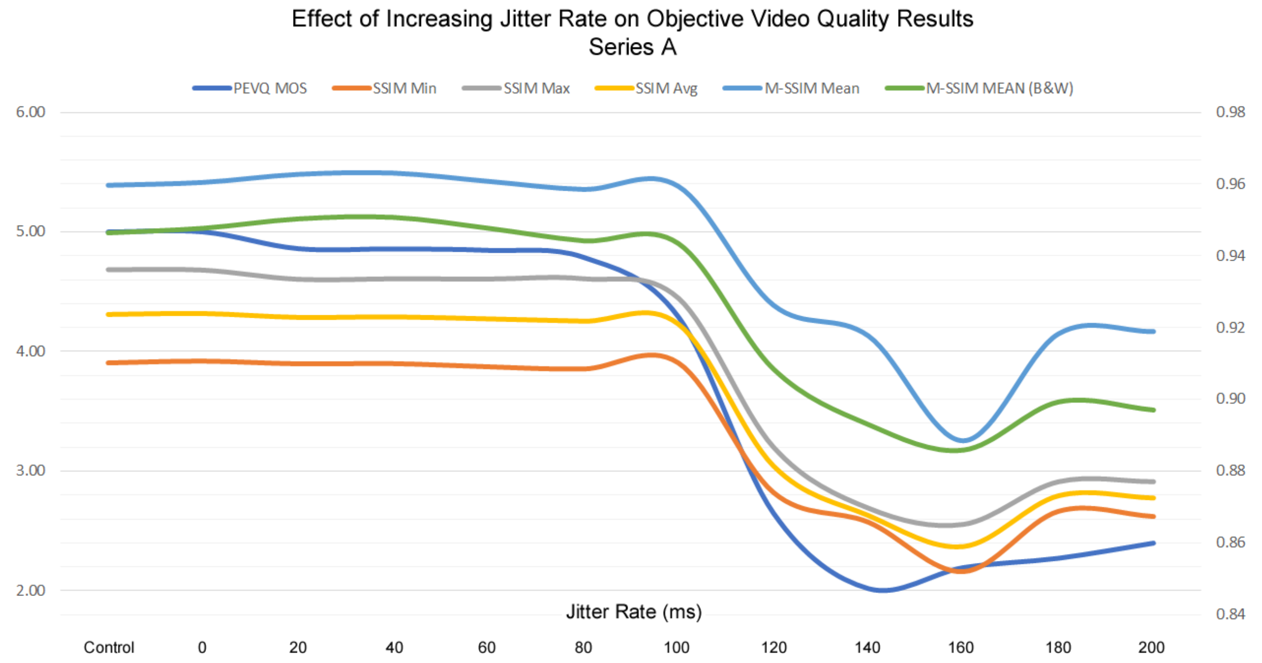


Figure 3: Unexpected Increase in Video Quality Scores Produced by Objective Tools as Jitter Rate Exceeded 140 ms

As described earlier, there were clips expressing 11 levels of jitter, 11 levels of packet loss, and one clip in each series with no jitter or packet loss, resulting in a total of 23 clips for each series set (A, F, G, H, I, and Z). For the results of each of the objective metrics over each of the video series, please see Appendix A – Measuring Audio and Video Quality.

A Pearson correlation analysis was performed among the 11 levels of jitter as well as the 11 levels of packet loss for each of the Objective Metrics and each of the Series. Table 2 and Table 3 show the correlation coefficients for the jitter and packet loss, respectively.

As can be seen in the jitter correlations (Table 2), there was a strong negative correlation for all of the data points (|coefficient| >.7), indicating that as the jitter increased, the Objective Measures decreased. Of the three Objective Measures, the PEVQ MOS had the strongest values, and PSNR and SSIM were both less so. There were only minor differences across the different Series.

Table 2: Correlation of Jitter vs. Objective Measure for Each Series

Jitter (ms)	PEVQ MOS	PSNR Avg	SSIM Avg
Series A	-0.90	-0.84	-0.88
Series F	-0.91	-0.86	-0.84
Series G	-0.92	-0.87	-0.90
Series H	-0.91	-0.91	-0.94
Series I	-0.92	-0.87	-0.78
Series Z	-0.86	-0.74	-0.76

The packet loss correlations (Table 3), were also negative, but the values, over all, were lower (some showed no correlation at all). Again, the PEVQ MOS had the strongest values, followed by the PSNR, and finally SSIM. In this case, there were some significant differences for the various Series; in particular, Series G had lower values for PEVQ MOS and PSNR.

Table 3: Correlation of Packet Loss vs. Objective Measure for Each Sequence

Packet Loss (%)	PEVQ MOS	PSNR Avg	SSIM Avg
Series A	-0.94	-0.64	0.00
Series F	-0.76	-0.72	-0.57
Series G	-0.38	-0.37	-0.50
Series H	-0.96	-0.67	-0.50
Series I	-0.99	-0.97	0.00
Series Z	-0.82	-0.67	-0.50

3.1.1 Discussion of Objective Assessment Results

Of the three Objective Measures, PEVQ had the strongest correlations with jitter and packet loss (with the exception of Series G). For jitter, both PSNR and SSIM performed well and it was difficult to rank them against each other, since results were series dependent. For packet loss, again PSNR and SSIM performed similarly, however in this case, poorly.

Jitter scores were consistently more correlated (negatively) with the jitter values, whereas the packet loss scores were less correlated with their respective values. This may be due to the innate self-corrections within the VTC system used.

For the jitter values and respective Objective scores, there was very little difference between the five series. For the packet loss values and respective Objective measures, Series F had slightly lower correlations and Series G had significantly lower correlations with the Objective scores.

3.2 VIDEO QUALITY – SUBJECTIVE ASSESSMENT

For all analyses, Microsoft Excel (version 14.0.6129.5000) with Analysis ToolPak add-in program was used. The Analysis ToolPak is an Excel add-in program that provides data analysis tools for financial, statistical, and engineering data analysis.

The raw data for subjective video quality (VQ) and perceived facial expression interpretability (FEI) ratings by trial by participant are given in Appendix E – Subjective Assessment Raw Data – VQ and FEI Ratings by Trial, by Participants, and Descriptive Statistics, as are the descriptive statistics (mean, standard deviation, min and max) for each trial. There was no missing data.

An F-Test was conducted to test for differences between mean VQ and FEI ratings, as shown in Table 4.

Table 4: F-Test: Two-Sample for Variances, VQ and FEI

	VQ	***FEI***
Mean	3.284076	3.709748
Variance	1.708419	1.218
Observations	115	115
df	114	114
F	1.402642	
P(F<=f) one-tail	0.036079	
F Critical one-tail	1.362605	

Since F (1.402642) is greater than F Critical one-tail (1.362605, p=0.05), the two variances are unequal. A t-Test to determine if the means are different was conducted, see Table 5.

Table 5: t-Test: Two-Sample Assuming Unequal Variances, VQ and FEI

	VQ	***FEI***
Mean	3.284076	3.709748
Variance	1.708419	1.218
Observations	115	115
Hypothesized Mean Difference	0	
df	222	
t Stat	-2.66843	
P(T<=t) one-tail	0.004091	
t Critical one-tail	1.651746	
P(T<=t) two-tail	0.008183	
t Critical two-tail	1.970707	

The FEI average rating (3.71) is statistically significantly greater than the VQ average rating (3.28).

Next, an F-Test was conducted to test for differences between mean jitter ratings and packet loss ratings as shown in Table 6.

Table 6: F-Test: Two-Sample for Variances, Jitter and Packet Loss

	Jitter	***Packet Loss***
Mean	3.253646	3.762294
Variance	2.234803	0.58019
Observations	120	110
df	119	109
F	3.851847	
P(F<=f) one-tail	3.3E-12	
F Critical one-tail	1.364678	

Since F (3.851847) is greater than F Critical one-tail (1.364678, p = 0.05), the two variances are unequal. A t-Test to determine if the means are different was conducted, see Table 7.

Table 7: t-Test: Two-Sample Assuming Unequal Variances, Jitter and Packet Loss

	Jitter	***Packet Loss***
Mean	3.253646	3.762294
Variance	2.234803	0.58019
Observations	120	110
Hypothesized Mean Difference	0	
df	180	
t Stat	-3.29032	
P(T<=t) one-tail	0.000602	
t Critical one-tail	1.653363	
P(T<=t) two-tail	0.001204	
t Critical two-tail	1.973231	

The packet loss average rating (3.76) is statistically significantly greater than the jitter average rating (3.25).

The results of an analysis of variance (ANOVA) run to test for differences among variances of the five video clips (A, F, G, H, I), are shown Table 8 and

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Table 9.

Table 8: Summary of Means and Variances for Video Clips (A, F, G, H, I)

Groups	Count	Sum	Average	Variance
a	46	174.2563	3.788179	1.334183
f	46	163.825	3.561413	1.43865
g	46	168.6813	3.666984	1.306426
h	46	145.7454	3.168379	1.622911
i	46	151.7819	3.299606	1.672286

Table 9: Single Factor ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	12.18012	4	3.045031	2.06458	0.086366	2.411768
Within Groups	331.8505	225	1.474891			
Total	344.0306	229				

Since F (2.06458) is not greater than F Critical (2.411768, p=0.05), the null hypothesis that the mean variances of all five video clips are equal cannot be rejected.

Finally, correlation coefficients were run on the mean ratings (VQ, FEI) and levels of jitter and packet loss, as shown in Table 10.

Table 10: Correlation Coefficients on the Mean Ratings (VQ, FEI) and Levels of Jitter and Packet Loss (Clips A, F, G, H, I)

Clip A: Jitter					**Clip A: Packet Loss**			
	Level	VQ	FEI			Level	VQ	FEI
Level	1.00				Level	1.00		
VQ	-0.89	1.00			VQ	-0.94	1.00	
FEI	-0.90	0.99	1.00		FEI	-0.93	0.97	1.00
Clip F: Jitter					**Clip F: Packet Loss**			
	Level	VQ	FEI			Level	VQ	FEI
Level	1.00				Level	1.00		
VQ	-0.92	1.00			VQ	-0.78	1.00	
FEI	-0.92	0.97	1.00		FEI	-0.82	0.94	1.00

Clip G: Jitter					Clip G: Packet Loss			
	Level	VQ	FEI			Level	VQ	FEI
Level	1.00				Level	1.00		
VQ	-0.92	1.00			VQ	-0.65	1.00	
FEI	-0.91	0.97	1.00		FEI	-0.51	0.96	1.00
Clip H: Jitter					**Clip H: Packet Loss**			
	Level	VQ	FEI			Level	VQ	FEI
Level	1.00				Level	1.00		
VQ	-0.89	1.00			VQ	-0.91	1.00	
FEI	-0.93	0.98	1.00		FEI	-0.83	0.91	1.00
Clip I: Jitter					**Clip I: Packet Loss**			
	Level	VQ	FEI			Level	VQ	FEI
Level	1.00				Level	1.00		
VQ	-0.88	1.00			VQ	-0.91	1.00	
FEI	-0.91	0.98	1.00		FEI	-0.96	0.95	1.00

Strong negative correlations were found between levels of jitter and packet loss and the corresponding VQ and FEI ratings. The negative correlation indicates that as the level of video distortion goes up, the VQ and FEI ratings go down. However, for Clip G, the correlation between packet loss level and VQ and FEI is not as strong. There are strong positive correlations between VQ and FEI ratings.

3.2.1 Discussion of Subjective Assessment Results

FEI ratings were consistently higher than VQ ratings (on average they were half a point higher on the Likert scale.) This suggests that even though participants noticed the jitter and packet loss distortions of the video, they were still able to adequately discern facial expressions of the subjects in the video. There was still a level of utility to the video clip, even though the video was noisy, at least up to a point. This might be a consideration when determining an acceptability threshold for automated video quality assessment tools. The FEI rating as a measure of utility seems to have merit and should be explored further.

Video clips that had packet loss distortions were rated consistently higher in terms of VQ and FEI than clips with jitter distortions (on average they were half a point higher on the Likert scale). This suggests that there is something more objectionable about jitter distortion on the subjective experience of video quality and the ability to interpret facial expressions. Since the experience of jitter affects the horizontal line displacement on the video, it makes sense that this would affect the ability to interpret facial expressions more than packet loss distortion due to network congestion. The source of the noise (jitter or packet loss) might be a consideration when determining an acceptability threshold for automated video quality assessment tools.

The study team found that there was no significant difference for the VQ or FEI mean ratings across the five different video clips. This suggests that the participants were not appreciably affected by background contrast, subject skin color, whether the subject wore glasses (potentially obscuring part of the face), or facial orientation when they gave their VQ and FEI ratings. Participants were able to rate VQ and FEI regardless of the content variations of the video that were presented. This human ability to rate video quality regardless of video content may not be found to the same degree in an automated tool set.

It was not surprising to observe a strong negative correlation between both VQ and FEI ratings and the levels of jitter and packet loss. However, the lack of strong correlation for Clip G for both VQ and FEI was somewhat surprising. An explanation could be that for Clip G, the subject focus was set at a greater standoff distance from the video camera than the other clips. It was also the only clip that showed the focus subject in a partial-face orientation. As a result, the size of the subject was smaller in terms of perceived visual angle and the actual video display resolution, and less of the face was visible to interpret facial expression. The smaller face area would cause distortion levels to have a greater impact on VQ and FEI ratings. This suggests that a court VTC configuration that shows a large standoff distance or the focus subject in profile would be less tolerant of network-induced distortion.

3.3 OBJECTIVE AND SUBJECTIVE MEASURES – CORRELATION

The goal of this portion of the study was to investigate how the objective ratings (PEVQ, PSNR, and SSIM) correlate to the mean subjective ratings (VQ, FEI) across all levels of jitter and packet loss. If any of the objective measures correlate highly with any of the subjective measures, this gives a basis for recommending that those objective assessment tools could be utilized in lieu of more time consuming and costly subjective testing to determine whether the video quality of a VTC network is acceptable for conducting court business.

This section gives a description of the combined measures data transformation, visualization of the data for comparison, and the correlation analysis.

Data derived from the Objective and Subjective data are included below. The raw data used for the Correlation between Objective and Subjective Assessment results can be found in their respective appendix. Correlation coefficients between objective and subjective measures for each individual video clip (both jitter and packet loss) can be found in Appendix A – Measuring Audio and Video Quality.

3.3.1 Data Analysis

For all combined correlation coefficient analyses, Microsoft Excel (version 14.0.6129.5000) with Analysis ToolPak add-in program was used. The Analysis ToolPak is an Excel add-in program that provides data analysis tools for financial, statistical and engineering data analysis. All correlation coefficients are Pearson r, where the linear correlation between two variables has a value between +1 and -1. On this scale +1 is a total positive correlation, 0 is no linear correlation, and -1 is a total negative correlation.

Prior to running the correlation coefficients, it was helpful to normalize the objective data to bring the values into a range more familiar in video image quality. In this case, the familiar range is the 1–5 point Likert scale, which is already in use for PEVQ (thus obviating the need for normalizing) and the two subjective measures (VQ and FEI). The objective data was transformed to the same scale as the subjective data with a uniform distribution using the “histeq” function in MATLAB and a linear remapping to the 1–5 Likert scale range.

The rationale and description of the data transformations that were accomplished for the objective data prior to the data visualizations and correlation analysis are included in the following section.

3.3.2 Histogram Equalization for Objective Data

When evaluating the data distributions of the objective measures (PEVQ, PSNR, and SSIM), the team found two major differences as compared to the subjective data (VQ and FEI). First, the range of the subjective data (as well as the PEVQ) followed the Likert scale (1–5), whereas the range of the SSIM and PSNR had different range values (17.63 to 31.75 and 0.7565 to 0.9239, respectively). In order to provide direct comparison of the objective measures to the subjective measures and also to apply meaningful scores across all measures, the team normalized SSIM and PSNR to the Likert scale of 1–5. Note that PEVQ scores are already normalized to the Likert scale.

The second difference was that the distribution of the objective data was skewed to the higher values, see Figure 4 for an example (SSIM), whereas the subjective data was more uniformly distributed, as shown in Figure 5 (VQ).

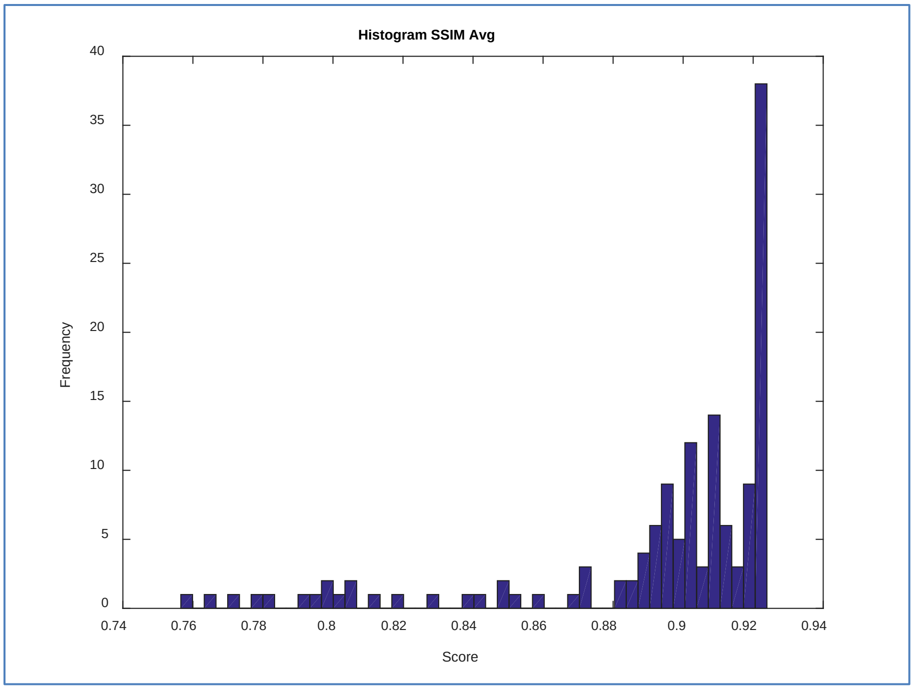


Figure 4: Histogram of SSIM Data
(Example of Skewed Objective Data)

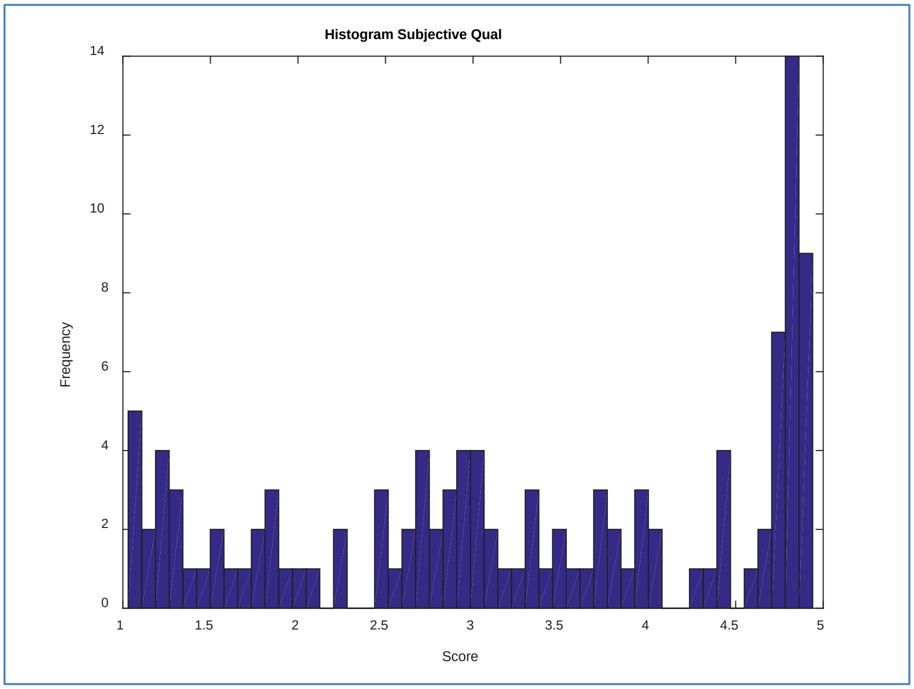


Figure 5: Histogram of the Subjective Video Quality Data
(Example of Uniformly Distributed Data)

Using the “histeq” function in MATLAB and a linear remapping to the 1–5 Likert scale range, the objective data was transformed to the same scale as the subjective data with a uniform distribution. Figure 6 shows the results for the SSIM data.

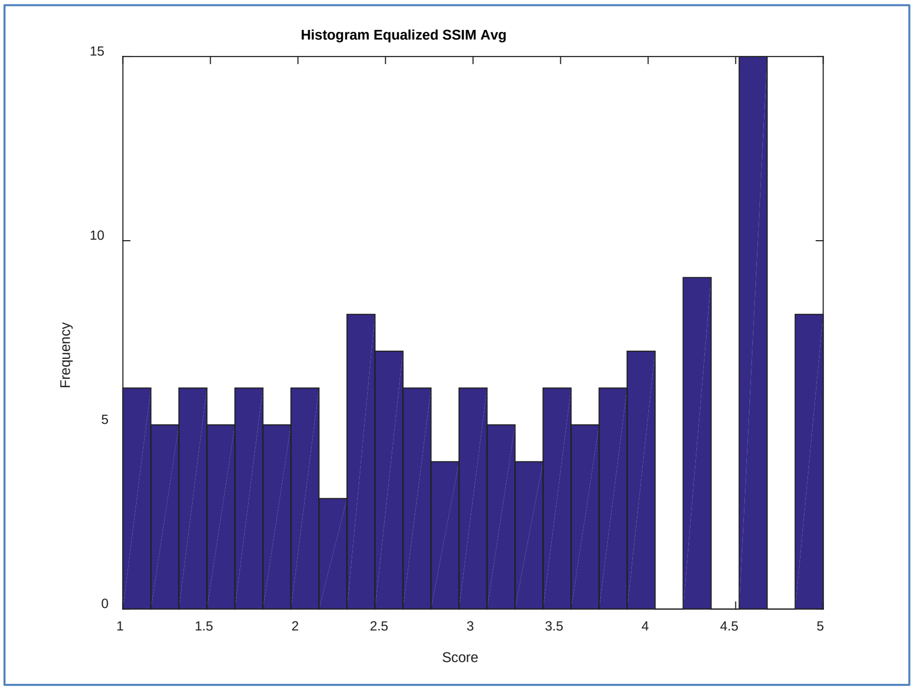


Figure 6: Histogram of SSIM Data After Histogram Equalization and Linear Remapping to 1–5 Range

3.4 VISUALIZATION OF THE NORMALIZED DATA FOR COMPARISON

It is helpful to look at the means of the subjective ratings (VQ and FEI) alongside the normalized objective ratings (PEVQ, PSNR, SSIM) to observe levels of agreement. Table 11 gives the color-coded legend and its relationship to the four Likert scale intervals. The series of tables from Table 12 through Table 23 have been color-coded to aid in visualization of agreement.

Each table is restricted to one video series and shows all jitter and packet loss levels as well as ‘none.’ Note that Series Z is not included because it was only used as a training series for the subjective data participants and not all clips were presented.

Following each color-coded table is a breakout table showing the level of agreement for the rating intervals for each clip in the series. For example, the number in the “0” column indicates the number of objective ratings that were in the **same** interval for the subjective ratings (e.g., green to green, yellow to yellow). The number in the +1 column indicates the number of objective ratings that were one interval **above** the subjective rating (e.g., green to yellow, orange to red). The number in the -2 column indicates the number of objective ratings that were two intervals **below** the subjective rating (e.g., red to yellow, orange to green). The closer the counts are to 0, the better agreement the objective ratings show to the subjective ratings when looking at an entire Likert interval.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Table 11: Likert-scale Interval Color-Code Legend

Color	Likert Definition Range	Likert Scale Range
Green	Good to Excellent	4.01 – 5.00
Yellow	Fair to Good	3.01 – 4.00
Orange	Poor to Fair	2.01 – 3.00
Red	Bad to Poor	1.00 – 2.00

Table 12: Comparison of Objective and Subjective Video Quality Ratings for Series A (all distortion levels)

Video Clip ID	Distortion Jitter (0–200 ms) Packet Loss (0–50%)	Objective Measures; Histogram Equalization			Subjective Measures (n=16)	
		PEVQ	PSNR	SSIM	VQ	FEI
A37	None	5.00	4.68	5.00	4.78	4.53
A51	0	5.00	4.68	5.00	4.88	4.66
A62	20	4.60	4.39	4.88	4.81	4.69
A50	40	4.60	4.39	4.88	4.81	4.63
A41	60	4.56	4.11	4.56	4.91	4.71
A72	80	4.35	4.11	4.23	4.81	4.63
A10	100	3.59	3.87	3.99	2.74	3.43
A80	120	1.85	2.86	1.73	1.91	2.39
A45	140	1.57	1.53	1.61	1.55	1.74
A13	160	1.61	1.97	1.57	1.31	1.74
A53	180	1.69	2.33	1.69	1.80	2.14
A28	200	1.73	2.17	1.65	1.83	2.16
A22	0	5.00	4.68	5.00	4.81	4.72
A63	5	4.19	3.87	4.23	4.84	4.69
A58	10	4.19	4.39	4.88	4.29	4.53
A38	15	4.07	3.87	4.23	4.59	4.56
A46	20	3.99	4.39	4.56	4.72	4.59
A89	25	3.95	4.11	4.56	4.03	4.43
A18	30	3.79	3.71	3.99	4.02	4.37
A34	35	3.67	4.11	4.56	3.63	4.28
A69	40	3.42	4.11	4.23	3.26	3.94

Video Clip ID	Distortion Jitter (0–200 ms) Packet Loss (0–50%)	Objective Measures; Histogram Equalization			Subjective Measures (n=16)	
		PEVQ	PSNR	SSIM	VQ	FEI
A42	45	3.51	3.59	3.79	3.44	4.16
A30	50	2.58	3.71	3.87	3.01	3.79

For Series A, the levels of agreement between subjective measures and objective measures by Likert interval are shown in Table 13.

Table 13: Number of Intervals of Agreement, Series A

Number of Intervals of Agreement	-3	-2	-1	0	+1	+2	+3
VQ							
PEVQ			4	18	1		
PSNR			3	14	6		
SSIM			1	19	3		
FEI							
PEVQ			9	14			
PSNR			4	18	1		
SSIM			6	17			

Table 14: Comparison of Objective and Subjective Video Quality Ratings for Series F (all distortion levels)

Video Clip ID	Distortion Jitter (0–200 ms) Packet Loss (0–50%)	Objective Measures; Histogram Equalization			Subjective Measures (n=16)	
		PEVQ	PSNR	SSIM	VQ	FEI
F88	None	4.72	3.38	4.56	4.84	4.63
F84	0	4.64	3.26	3.99	4.89	4.75
F29	20	4.07	3.26	4.23	4.44	4.74
F69	40	4.19	3.26	4.23	4.81	4.63
F41	60	4.27	3.26	4.56	4.91	4.75
F32	80	4.19	3.26	4.56	4.93	4.75
F89	100	2.45	2.86	2.94	3.13	4.13
F98	120	2.05	2.49	3.06	2.27	3.27
F20	140	1.12	1.24	1.12	1.11	1.64
F93	160	1.53	1.32	1.48	1.51	1.98
F68	180	1.44	1.24	1.36	1.38	1.76
F54	200	1.28	1.04	1.04	1.09	1.39
F16	0	4.47	3.38	4.56	4.80	4.81
F38	5	2.86	2.98	3.59	2.89	3.99
F57	10	3.99	3.26	4.23	3.78	4.42
F65	15	3.87	3.26	4.23	3.76	4.28
F12	20	3.83	3.10	3.99	3.76	4.49
F10	25	3.75	3.06	3.87	3.61	4.33
F53	30	3.51	3.02	3.75	3.19	4.23
F60	35	3.22	3.06	3.75	2.96	3.87
F76	40	2.37	2.86	3.42	2.86	3.63
F39	45	2.78	2.94	3.63	2.74	3.62
F56	50	2.45	2.94	3.59	2.46	3.62

For Series F, the levels of agreement between subjective measures and objective measures by Likert interval are shown in Table 15.

Table 15: Number of Intervals of Agreement, Series F

Number of Intervals of Agreement	-3	-2	-1	0	+1	+2	+3
VQ							
PEVQ			1	21	1		
PSNR			9	13	1		
SSIM			8	13	2		
FEI							
PEVQ		1	10	12			
PSNR		1	17	5			
SSIM		1	4	18			

Table 16: Comparison of Objective and Subjective Video Quality Ratings for Series G (all distortion levels)

Video Clip ID	Distortion Jitter (0–200 ms) Packet Loss (0–50%)	Objective Measures; Histogram Equalization			Subjective Measures (n=16)	
		PEVQ	PSNR	SSIM	VQ	FEI
G26	None	3.26	5.00	2.90	4.72	4.47
G98	0	3.18	4.96	2.86	4.91	4.58
G32	20	3.10	4.92	2.37	4.87	4.62
G74	40	3.02	4.84	2.29	4.78	4.58
G50	60	3.02	4.84	2.29	4.75	4.58
G41	80	2.94	4.84	2.29	4.81	4.55
G46	100	2.33	4.11	2.01	3.06	4.08
G94	120	1.57	2.70	1.44	1.84	2.46
G24	140	1.65	1.40	1.24	1.74	2.41
G86	160	1.12	1.36	1.12	1.31	1.52
G55	180	1.44	1.44	1.32	1.72	2.35
G75	200	1.32	1.57	1.20	1.34	1.64
G29	0	3.26	5.00	2.90	4.84	4.70
G37	5	2.09	3.59	1.77	2.78	3.55
G54	10	2.90	4.84	2.17	4.78	4.64
G73	15	2.86	4.84	2.09	4.43	4.54

Video Clip ID	Distortion Jitter (0–200 ms) Packet Loss (0–50%)	Objective Measures; Histogram Equalization			Subjective Measures (n=16)	
		PEVQ	PSNR	SSIM	VQ	FEI
G48	20	2.82	4.68	2.09	3.96	4.36
G33	25	2.74	4.68	2.01	3.78	4.44
G81	30	2.70	4.68	2.01	3.91	4.28
G70	35	2.66	4.52	1.89	3.31	4.07
G69	40	2.54	4.52	1.89	3.31	3.99
G96	45	2.54	4.52	1.89	2.93	3.96
G45	50	2.21	3.87	1.85	2.79	3.66

For Series G, the levels of agreement between subjective measures and objective measures by Likert interval are shown in Table 17.

Table 17: Number of Intervals of Agreement, Series G

Number of Intervals of Agreement	-3	-2	-1	0	+1	+2	+3
VQ							
PEVQ		3	12	8			
PSNR				13	9	1	
SSIM		11	7	5			
FEI							
PEVQ		8	13	2			
PSNR			3	18	2		
SSIM		17	4	2			

Table 18: Comparison of Objective and Subjective Video Quality Ratings for Series H (all distortion levels)

Video Clip ID	Distortion Jitter (0–200 ms) Packet Loss (0–50%)	Objective Measures; Histogram Equalization			Subjective Measures (n=16)	
		PEVQ	PSNR	SSIM	VQ	FEI
H40	None	4.76	2.82	3.30	4.47	4.25
H77	0	4.72	2.49	3.06	4.47	4.63
H67	20	4.84	2.78	3.30	4.84	4.63
H58	40	3.71	2.62	2.98	3.96	4.53
H52	60	2.25	2.29	2.49	2.91	3.78
H38	80	4.84	2.78	3.30	4.81	4.72
H66	100	2.01	1.85	2.13	1.99	2.89
H98	120	1.28	1.48	1.53	1.25	1.83
H26	140	1.12	1.12	1.40	1.24	1.63
H37	160	1.12	1.20	1.28	1.09	1.69
H97	180	1.12	1.04	1.16	1.03	1.25
H48	200	1.12	1.00	1.04	1.05	1.22
H12	0	4.72	2.58	3.18	4.84	4.69
H54	5	4.39	2.70	3.42	3.73	4.36
H95	10	4.11	2.70	3.42	3.48	4.32
H88	15	3.71	2.58	3.30	3.31	4.26
H80	20	3.63	2.49	3.18	2.86	3.78
H69	25	3.34	2.33	3.06	3.00	4.05
H89	30	2.90	2.49	3.18	2.97	4.02
H71	35	2.66	2.41	3.18	2.54	3.38
H53	40	2.58	2.41	3.18	2.63	3.63
H86	45	2.25	2.33	3.06	2.63	3.69
H49	50	1.89	1.69	2.58	1.47	1.99

For Series H, the levels of agreement between subjective measures and objective measures by Likert interval are shown in Table 19.

Table 19: Number of Intervals of Agreement, Series H

Number of Intervals of Agreement	-3	-2	-1	0	+1	+2	+3
VQ							
PEVQ				18	5		
PSNR		5	4	14			
SSIM			6	9	8		
FEI							
PEVQ		1	7	15			
PSNR		11	6	6			
SSIM		1	11	10	1		

Table 20: Comparison of Objective and Subjective Video Quality Ratings for Series I (all distortion levels)

Video Clip ID	Distortion Jitter (0–200 ms) Packet Loss (0–50%)	Objective Measures; Histogram Equalization			Subjective Measures (n=16)	
		PEVQ	PSNR	SSIM	VQ	FEI
I28	None	3.42	1.97	2.41	4.83	4.64
I39	0	3.63	1.73	2.01	4.78	4.75
I25	20	3.10	2.25	2.86	4.69	4.69
I38	40	2.54	1.97	2.58	3.96	4.42
I55	60	3.18	2.17	2.74	4.91	4.78
I45	80	3.02	2.17	2.58	4.69	4.72
I89	100	1.77	1.65	2.29	1.89	3.06
I87	120	1.36	1.28	1.24	1.18	1.86
I47	140	1.53	1.40	1.77	1.61	2.21
I72	160	1.12	1.08	1.00	1.05	1.13
I59	180	1.12	1.16	1.36	1.20	1.59
I49	200	1.12	1.12	1.08	1.20	1.51
I80	0	3.14	2.25	2.74	4.94	4.75
I34	5	2.78	2.09	2.74	4.34	4.59
I67	10	2.37	2.09	2.74	3.53	4.16

Video Clip ID	Distortion Jitter (0–200 ms) Packet Loss (0–50%)	Objective Measures; Histogram Equalization			Subjective Measures (n=16)	
		PEVQ	PSNR	SSIM	VQ	FEI
I37	15	2.25	2.09	2.74	3.09	4.10
I23	20	2.13	1.93	2.74	3.03	4.00
I69	25	2.09	1.89	2.58	2.70	3.84
I78	30	1.97	1.89	2.49	2.46	3.70
I42	35	1.97	1.85	2.45	2.51	3.65
I70	40	1.89	1.77	2.41	2.68	3.74
I68	45	1.81	1.65	2.17	2.21	3.33
I62	50	1.77	1.61	2.29	2.07	3.01

For Series I, the levels of agreement between subjective measures and objective measures by Likert interval are shown in Table 21.

Table 21: Number of Intervals of Agreement, Series I

Number of Intervals of Agreement	-3	-2	-1	0	+1	+2	+3
VQ							
PEVQ		3	13	7			
PSNR	2	7	8	6			
SSIM		7	4	11	1		
FEI							
PEVQ		11	8	4			
PSNR	4	14	1	4			
SSIM		8	11	4			

To examine whether the objective ratings (PEVQ, PSNR, and SSIM) correlate to the mean subjective ratings (VQ, FEI) across all levels of jitter and packet loss, the following analyses were conducted and results obtained.

The number of objective rating intervals that were in perfect agreement with the subjective scores across all Series was collected, as shown in Table 22.

Table 22: Number of Objective Rating Intervals in Perfect Agreement with Subjective Rating Intervals Across All Series

	Series A	Series F	Series G	Series H	Series I	Total
VQ						
PEVQ	18	21	8	18	7	72
PSNR	14	13	13	14	6	60
SSIM	19	13	5	9	11	57
Total	**51**	**47**	**26**	**41**	**24**	**189**
FEI						
PEVQ	14	12	2	15	4	47
PSNR	18	5	18	6	4	51
SSIM	17	18	2	10	4	51
Total	**49**	**35**	**22**	**31**	**12**	**149**

The highest agreement between objective and subjective VQ ratings occurred for Series A, followed by F and H. There was similar lower agreement scores for Series G and I. For FEI, Series A was again highest, followed by F and H. However, Series I was significantly lower than Series G.

The number of objective rating intervals that were in perfect agreement with the subjective scores across all Series was collected, as shown in Table 22, suggests that video content has an effect on the agreement between Objective and Subjective measures of video quality, and that the effect is a more pronounced when Objective video quality was compared to Subjective measures of FEI. This indicates that when best practices are not followed during the transmission of the video hearing, the ability of the viewer to discern facial expressions and thus interpret demeanor, can be impacted, despite objective video quality measures indicating that the video signal is acceptable.

When looking at the three objective measures and how many rating intervals were in perfect agreement over the two subjective measures combined, refer to Table 23. All three objective tools displayed better agreement for subjective scores for VQ than for FEI. The scores for three objective measures and how many rating intervals were in perfect agreement over the two subjective measures combined, (Table 23) suggest that of the three objective tools tested, PEVQ is more likely to align with subjective scores for video quality, than are PSNR and SSIM. Conversely, the scores for FEI showed slightly better alignment with the scores from PSNR and SSIM. This suggests that while PEVQ could be a better objective tool to predict subjective scores for video quality, PEVQ is not most likely to predict subjective FEI.

Table 23: Objective Rating Intervals in Perfect Agreement over the Subjective Measures Combined

Objective Measure	VQ	FEI	Combined
PEVQ	72	47	119
PSNR	60	51	111
SSIM	57	51	108
Total	**189**	**149**	**338**

As in the visualization tables, Series Z was not used in this analysis because it was used as a partial training set for participants. Table 24 gives the color-code legend and its trace to the three widely accepted intervals for determining correlation 'goodness'. The following series of tables (Table 25 through Table 27) has been color-coded to aid in visualization of correlation. The highest level of objective correlation with each subjective VQ and FEI are marked with an asterisk (*).

Table 24: Correlation Color-Code Legend

Color	Correlation Coefficient Definition Range	Correlation Coefficient Range
White	Strong Correlation	.71–1.00
Aqua	Moderately Strong Correlation	.50–.70
Peach	Weak Correlation	Less than .50

Table 25: Correlation Coefficients for All Measures across All Clips and All Distortion Levels

Combined Measures	PEVQ MOS	PSNR Avg	SSIM Avg
Subj VQ	0.87*	0.71	0.69
Subj FEI	0.84*	0.70	0.73

Table 26. Correlation Coefficients for All Measures across All Clips and All Jitter Levels

Jitter	PEVQ MOS	PSNR Avg	SSIM Avg
Subj VQ	0.91*	0.77	0.80
Subj FEI	0.89*	0.75	0.80

Table 27: Correlation Coefficients for All Measures across All Clips and All Packet Loss Levels

Packet Loss	PEVQ MOS	PSNR Avg	SSIM Avg
Subj VQ	0.72*	0.58	0.36
Subj FEI	0.71*	0.52	0.37

4. OVERALL CONCLUSIONS

Video content has an effect on the agreement between Objective and Subjective measures of video quality, and that the effect is a more pronounced when Objective video quality was compared to Subjective measures of FEI. This indicates that when best practices are not followed during the transmission of the video hearing, the ability of the viewer to discern facial expressions and thus interpret demeanor, can be impacted, despite objective video quality measures indicating that the video signal is acceptable.

Of the three objective tools tested, PEVQ is more likely to align with subjective scores for video quality, than are PSNR and SSIM. Conversely, the scores for FEI showed slightly better alignment with the scores from PSNR and SSIM. This suggests that while PEVQ could be a better objective tool to predict subjective scores for video quality, PEVQ is not most likely to predict subjective FEI.

4.1 SUPPORTING CONCLUSIONS

The RT&E Center’s testing was intended to investigate quality issues resulting from network transmission and bandwidth, with particular interest in the ability of VTC to allow users to observe aspects of demeanor such as nonverbal cues, facial expressions and body language which were noted as critical to determining whether a VTC hearing is experienced similarly to an in-person hearing.

This study examined the effects of file degradation typical of network issues specifically to evaluate the usefulness of objective and subjective methods of measuring video quality and the impact of packet loss and jitter (latency) on the perceived effectiveness of VTC hearings.

The overall objective of the RT&E Center research was to conduct experimental quantitative and qualitative research to determine whether quantitative metrics for video quality could be found that correlate highly with human subjects’ perceived video quality.

The RT&E Center, in consultation with NIJ, scoped the study using the objectives as follows:

- Continue to strengthen the foundation established in Davis et al.[24]
- Identify and assess methods, metrics, and tools for measuring video quality;
- Evaluate the effects of network degradation on video quality through objective and subjective testing;
- Demonstrate a methodology that can show whether the video is ‘sufficient’ for a VTC hearing;
- Attempt to identify a repeatable, cost effective method to validate the perceived quality of a VTC system; and
- Develop guidance on use of VTC in pretrial hearings.

The study results are presented with the caveat that the small sample size of both human subjects and video exemplars, and the lack of diversity among the video clips selected do not support broad conclusions about the entire sample space. These conditions were mandated by cost and schedule constraints, and by the relative lack of fundamental research results on videoconferencing in courtrooms.

Three quantitative metrics were selected to objectively evaluate the quality of sample videos: PEVQ, and Structural SIMilarity (SSIM), which are based on human perception models; and Peak Signal-to-Noise Ratio (PSNR) which is not human perception based. Then using a five-point Likert-type scale aligned to ITU-T MOS definitions, human subjects rated their perceptions of the same video clips for subjective video quality (VQ) corresponding to bad (1), poor (2), fair (3), good (4), excellent (5). As a model of how video quality impacts a viewer’s ability to identify what is happening, assess emotions and motivations of the speaker, and derive any additional information that would enable them to make judgments regarding the trustworthiness of the speaker and the veracity of his or her testimony, participants were also asked to rate their ability to interpret facial expressions in the video clips, using a scale ranging from ‘cannot at all interpret’ (1) through ‘can very easily interpret’ (5) the facial expression.

Although FEI ratings were consistently about half a point higher on the Likert scale higher than VQ ratings, the study found strong positive correlations between VQ and FEI ratings. This suggests that while participants noticed the jitter and packet loss distortions of the video, they believed they were still able to adequately discern facial expressions of the subjects in the video. Thus, there was utility in videos with some levels of apparent noise. Therefore, FEI rating as a measure of utility should be explored further to determine an acceptability threshold for automated video quality assessment tools. It should be noted that the short duration of the tested video clips did not allow evaluation of fatigue or frustration, which might cause a viewer to reject the utility of a degraded video after some length of time.

The source of the noise (jitter or packet loss) might be a consideration when determining an acceptability threshold for automated video quality assessment tools. Video clips that had packet loss distortions were rated consistently higher in terms of VQ and FEI than clips with jitter

[24] Davis et al., “Research on Videoconferencing.”

distortions (on average they were half a point higher on the Likert scale). This could suggest that there is something more objectionable about jitter distortion, which caused horizontal line displacement on the video, on the subjective experience of video quality and the ability to interpret facial expressions. However, it is possible that the equipment used in the experimental test bed was better able to compensate for packet loss and that the full visual effect of the degradation was not captured in the test clips viewed by the participants.

Distortion levels appeared to have greater impact on VQ and FEI ratings of the test video with smaller faces. This suggests that a court VTC configuration that shows a large standoff distance or displays the focal subject in profile would be less tolerant of network-induced distortion. Participants were able to rate VQ and FEI regardless of the other content variations of the video that were presented, with no significant difference for the VQ or FEI mean ratings across the five different video clips. This human ability to rate video quality regardless of video content may not be found to the same degree in an automated tool set.

Since the focus of this subjective study was to generate benchmark ratings for the objective study, there was no planned in-depth analysis of the participant data. It should be considered investigational in nature as no results are definitive with the limited participant sample size of 16 and the limited pool of videos. In future work, it would be interesting to explore if there are differences in video quality ratings based on gender of the participant or the apparent race-matching of the participant to the video subject. Additionally, a study into the effect of experience in video quality assessment should be conducted. Future investigations should systematically select a diverse range of video clips that are more closely aligned with the court VTC setting.

For each video series, the objective tools indicated that video quality was relatively steady for jitter rates of 0 through approximately 80 ms, whereupon video quality scores decreased. All tools then showed an increase in video quality scores around 160 ms of jitter suggesting improved video quality, then a second drop in quality. The team concluded that this trend resulted from the VTC system compensating for reduced bandwidth, leading to an increase in video quality scores. Video quality scores for packet loss scores were generally much steadier, making any pattern difficult to distinguish. Future studies should utilize methods to ensure that the intended distortion levels are accurately captured in the test video recordings.

For the jitter values and respective objective scores, there was very little difference between the five clips. For the packet loss values and respective Objective measures, one video series had significantly lower correlations with the Objective scores and another had slightly lower correlations.

Jitter scores were consistently more negatively correlated with the level of introduced jitter, whereas the packet loss scores were less correlated with their respective levels. This may be due to the innate self-corrections within the VTC system used.

Of the three objective measures, PEVQ scores had the strongest correlations with the levels of introduced jitter and packet loss. For jitter, both PSNR and SSIM performed well but sequence dependence made it difficult to rank them against each other. Both PSNR and SSIM scores indicated poor correlation to the introduced levels of packet loss.

The correlation between the objective results and the subjective results are of primary interest for this study. Prior to calculating the correlation coefficients, the objective data was normalized to align with the Likert intervals used in the Subjective ratings. Objective tools measurements of video quality were found to correspond to the amount of introduced quality degradation due to jitter, but those objective values cannot indicate the impact on human perception. If any of the objective measures correlate highly with any of the subjective measures, this gives a basis for recommending that those objective assessment tools could be utilized to determine the adequacy of video quality of a VTC network for conducting court business.

When visualizing the data normalized across the Likert intervals, PEVQ was most closely aligned with the interval ratings of the subjective VQ measure, followed by PSNR and SSIM in that order. The interval rating alignment for FEI showed all three objective measures closely clustered with much lower alignment than with VQ. Objective measures of video quality would be expected to align more closely with subjective measures of video quality than with subjective measures of utility, which the objective tools are unable to measure. Lower correlation between FEI and objective video quality is consistent with the finding that participants rated their ability to interpret facial expression consistently higher than overall video quality.

It should also be noted that PEVQ had the lowest score for FEI, with the trend for the PEVQ rating to be one interval lower (-1) than the subjective rating intervals. On the other hand, PSNR and SSIM tended to have intervals spread across the + 2 range. The results suggested that video context impact objective and subjective quality scores, though this was not specifically tested in the visualization of the normalized data.

Though not the main thrust of the combined measures analysis, the visualization of normalized data approach appears to have some merit. While more study is needed, objective measures appear to be more conservative than human participants in the scoring of video. As a result, if the objective measure determines that video is bad or poor, there is a high likelihood the video is not usable. If the objective measure determines the video is fair, good or excellent – then there is a high likelihood the video is usable.

With regard to the correlation coefficient analysis, when jitter and packet loss are taken together, there are strong positive correlations between both subjective measures and PEVQ and PSNR. PEVQ is noticeably more highly correlated to the subjective measures than PSNR or SSIM. For SSIM, there is a strong correlation to FEI, but only moderately strong for VQ.

There was strong positive correlation between all three objective measures and both subjective measures for jitter, with PEVQ most strongly correlated with subjective measures.

There was lower positive correlation between PEVQ and PSNR and the two subjective measures for packet loss. PEVQ is the most strongly correlated with subjective measures; PSNR showed moderately strong correlation with the subjective measures, while SSIM showed only a weak correlation to the subjective measures for packet loss.

That PEVQ is most highly correlated overall with both subjective measures (VQ and FEI) is not surprising, since PEVQ utilizes a human model in its algorithm. For jitter only, the higher overall correlation coefficients suggest that jitter is more of a factor when participants give their video quality scores and when the three objective measures calculate their scores. For packet loss only, PEVQ bears a borderline strong correlation to VQ and FEI, while PSNR is barely moderately strong and SSIM is a weak correlation. This suggests that packet loss is less proscriptive for both the human participant scoring of video quality as well as the objective scoring calculations. But it is also possible that the test equipment was better able to compensate for packet loss than jitter.

Based on the combined objective and subjective measures analysis, all three objective tools appear to be acceptable alternatives to subjective measures. The rank order of goodness-of-fit for the objective tools are:

1. PEVQ
2. PSNR
3. SSIM

4.2 FURTHER STUDY

Since the focus of this subjective study was to generate benchmark ratings for the objective study, there was no planned in-depth analysis of the participant data. It should be considered investigational in nature as no results are definitive with the limited participant sample size of 16 and the limited pool of videos. In future work, it would be interesting to explore if there are differences in video quality ratings based on gender of the participant or the apparent race-matching of the participant to the video subject. Additionally, a study into the effect of experience in video quality assessment should be conducted.

Since the subjective study was conjoined with the objective study, there was a predetermined data-set of video clips (the same ones used in the objective study). While there was diversity of background contrast, some apparent race / ethnicity diversity, and the use of glasses, the videos themselves were not representative of a court VTC setting, Table 1. Future investigations should systematically select a diverse range of video clips that are more closely aligned with the court VTC setting.

For the jitter values and respective Objective scores, there was very little difference between the five clips. Retest using simulated network test bed that does not include auto-corrections for video degradation.

The results for Series G which is the only series which shows the subject from the side and at increased apparent distance from the viewer, suggested that the distance and angle between the camera and the subject may affect the impact of network distortion on both objective and subjective measurements of video quality. This may be because any loss or corruption of the transmitted video data represents a larger proportion of the number of pixels available to convey any particular aspect of the subject, such as facial features.

Participants were able to rate VQ and FEI regardless of the content variations of the video that were presented. This suggests that the participants were not appreciably affected by background contrast, subject skin color, whether the subject wore glasses (potentially obscuring part of the face), or facial orientation when they gave their VQ and FEI ratings.

Strong negative correlations were found between levels of jitter and packet loss and the corresponding VQ and FEI ratings indicating that the participants were able to recognize a loss in video quality related to increasing distortion of the video clips. FEI rating is significantly higher than the corresponding VQ rating indicating that even though participants noticed the jitter and packet loss distortions of the video, they were still able to adequately discern facial expressions of the subjects in the video. The FEI rating as a measure of utility seems to have merit and should be explored further.

However, for Clip G, the correlation between packet loss level and VQ and FEI is not as strong. There are strong positive correlations between VQ and FEI ratings.

Video clips that had packet loss distortions were rated consistently higher in terms of VQ and FEI than clips with jitter distortions (on average they were half a point higher on the Likert scale). However, when combined with the Objective data the difference may be related to the test bed equipment more effectively correcting for packet loss than jitter.

Although the subjective scores suggested that viewers found that jitter distortion had a more noticeably negative impact on both video quality and the ability to interpret facial expressions, additional study is needed to determine if the effect noted in subjective testing was due to higher sensitivity of participants to jitter, or due to the test bed being more efficient at self-correcting for packet loss.

Study results suggested that humans are able to rate video quality regardless of video content or filming conditions, while video quality scores generated by objective tools are more content dependent. Additional study is needed to validate this apparent difference.

Results of objective testing indicated that the subject had a stronger impact on objective video quality scores than it did on subjective scores. Further study is needed to verify that this is not found to the same degree in an automated tool set.

APPENDIX A. MEASURING AUDIO AND VIDEO QUALITY

Appendix A – List of Figures

A.1 INTRODUCTION

In measuring either the audio or video quality of a videoconferencing stream, there are a number of basic requirements for any measurement methodology.[1]

- Objectivity: Results must be reproducible for a range of observers/listeners.
- Reliability: Results must be repeatable for a single listener.
- Validity: Results must measure the desired audio and video characteristics.
- Sensitivity: Results must achieve a level of granularity commensurate with those of a listener.
- Comparability: Results must apply to a wide range of perceived qualities and support comparisons between groups and conditions.
- Utility: Results must provide useful information.

A.2 OBJECTIVE VS. SUBJECTIVE MEASUREMENTS

There are two basic types of measurements made for both audio and video: objective and subjective. Objective measurements tend to be performed at the signal level. Differences in the input and output signals are measured, thus objective measurements measure the performance of the delivery system. In contrast, subjective measurements require subjective assessments from human observers. Subjective measurements are qualitative and tend to measure the quality of the product.

It is generally accepted that the most accurate measurements of audio or video quality require evaluation from human subjects.[2] Because perception and biases differ from one person to the next, deriving statistically valid conclusions can be challenging. International Telecommunications Union (ITU) Telecommunication Standardization Sector (ITU-T) standards recommend that subjective evaluations utilize at least four observers, with 10–15 observers preferred. The presence of 10–15 trained observers assessing the quality of a videoconferencing stream in a courtroom is likely to be disruptive as well as prohibitively expensive. In addition, concerns about the safety and welfare of human participants involved in behavioral research and the need to maintain privacy and confidentiality of some proceedings further discourage the options of conducting routine subjective quality assessments of courtroom Video Teleconferencing (VTC) systems.

Objective measurements of video or audio quality usually derive a quality measurement based upon characteristics of the input and output signals, although some tools incorporate models of human perception. Also, because they are automated, they do not require large numbers of trained observers to sit and watch video and assess its quality for hours on end. As a result, objective measurements are in general repeatable and they provide quantitative results that support comparison. Automated objective measurement techniques are thus less intrusive and less costly than subjective measurements in a courtroom setting.

[1] Côtė, N., “Integral and Diagnostic.”

[2] Huynh-Thu et al. “Study of Rating Scales.”

However, objective measures of audio video quality have their own limitations. First, the level of packet loss or delay in a signal may not have a predictable effect on a human observer. Two signals with equal levels of distortion can have significantly different value to human observers, depending on which packets are distorted and the distribution of distortion within an image. Thus, assigning meaning to an objective measure of quality to fit a large number of varying settings and use cases can be challenging. Second, the objective assessment of audio and video quality of the delivered signal through a simulated packet data network may not capture the effects of factors beyond the network. The audio and video captured in environments with insufficient lighting, poor acoustical pickup, or excessive background noise may impact the objective measurements and may be of limited value due to poor video quality caused by environmental factors and audio/video equipment limitations.

A.3 TYPES OF OBJECTIVE MEASUREMENT

For purposes of this study, the RT&E Center chose three automated tools for measurement of video quality. Two of these tools, Structural SIMilarity (SSIM) and Perceptual Evaluation of Video Quality (PEVQ) attempt to measure video quality in a manner consistent with human perception. The third, Peak Signal-to-Noise Ratio (PSNR), provides a simple, but readily repeatable and well understood metric, based upon measureable network metrics, without a human visual model.

In addition, a discussion of Perceptual Evaluation of Speech Quality (PESQ), currently the most commonly used standard for measuring audio quality in telephony systems, has been included in this appendix. Because audio quality measurement has been taking place for decades, in association with telephony, and is therefore more mature than video quality measurement, and because existing videoconferencing systems prioritize audio over video (these systems will allow video to degrade or even be discontinued, before allowing audio degradation), no laboratory work was performed to assess audio measurement techniques.

The most commonly accepted metric for measuring either audio or video quality is the mean opinion score (MOS). MOS was adopted by ITU-T as a way of quantifying the perceived quality of media [also referred to as the Quality of Experience (QoE)]. MOS consists of a five-point scale (1 = bad; 2 = poor; 3 = fair; 4 = good; 5 = excellent). Observers are asked to assign a QoE value to audio or video media, and the arithmetic mean is computed. While MOS was originally developed as a metric for use in subjective evaluations, there are tools that perform objective measurements and present results as MOS.

A.3.1 Full Reference vs. No Reference

Another way of classifying video or audio quality measurement methodologies is by whether or not they use a reference file:

- Full Reference: Full reference quality metrics assess quality by comparing images from the output video file to an input reference file. Full reference quality metrics use clearly defined mathematical algorithms to create readily repeatable quantitative output. Results of these measurements lend themselves to consistent, if not necessarily relevant, interpretation. Because these measurement tools require a reference file be recorded, they do not reflect every factor that contributes to video quality. For example, full reference assessment in the instance of video conferencing system being utilized in the courtroom, a measurement tool will provide an accurate objective measurement of the loss of information as images are passed from codec to codec across a network, but they do not reflect losses of information at the camera or due to lighting and other environmental effects at the input site. Thus, they measure the performance of the delivery system, but not the quality of the product.
- Partial Reference: Partial reference video quality metrics extract features from the input video for comparison to the output video. They are essentially a simplified, more efficient but less comprehensive, version of full reference video quality measurement.
- No Reference: As the name implies, no reference video quality metrics do not rely on the use of an input file as a reference file. Subjective video quality measurements can be either no reference or full reference, depending on the desired result. Specifically, if the goal is to assess the quality or utility of a specific video, the output file can be presented to observers without an input reference file. On the other hand, if the goal is to assess the degradation across a network, observers can be asked to compare input files to output files.

A.4 MEASURING AUDIO QUALITY

As with video, the preferred method for measuring the quality of audio is to perform subjective tests with a sufficient number of trained observers. ITU-T has developed recommendations for computing MOS for audio quality. However, as is the case with video quality metrics, field measurements of audio quality with human subjects are expensive and impractical, and results are not readily repeatable.

For these reasons, tools for achieving objective measures of audio quality have also been developed. However, as in the case with video quality assessment tools, these metrics may provide measurements that are accurate but not useful. Specifically, they may not effectively measure the characteristics of sound or voice communications that most affect perception.

Perceptual models, which distinguish audible sound distortions from inaudible sound distortions, have been in use since the 1980s. By the 1990s, these models had been enhanced by the recognition that not only the quantity but also the distribution of audio distortion can affect perception. In 1996, ITU-T adopted the Perceptual Speech Quality Measure (PSQM) model of audio quality as a recommendation (P.861). However, because PSQM was designed to measure quality between voice codecs and did not anticipate advances such as Voice over Internet Protocol (VOIP), it did not correlate well with subjective assessments of certain types of commonly occurring network distortions. Specifically, in VOIP applications, delays between the output and reference files are not always constant. PESQ, which was designed to correct some of the deficiencies of PSQM, was adopted as a recommendation (P.862). Like PEVQ (described in the previous section), PESQ expresses quality measurements as MOS.

PESQ combines an upgraded version of the audio-cognitive model used in PSQM with a time alignment algorithm to enable it to compensate for the variable delay in Internet Protocol (IP) transmissions. An input reference signal and an impaired output signal are both input to PESQ, which correlates features to temporally align the two signals. A key feature of the algorithm is its ability to estimate the confidence in having one or more than one delay in an interval to distinguish start and stop times of speech and pauses. The algorithm is capable of resolving delay variances during both speech and silence. Aligned files are then input into a perceptual model that transforms the two files into forms analogous with features of human speech. Specifically, PESQ transforms the aligned input and output files from a time-amplitude domain to a frequency-loudness domain. The difference between the two resulting signal representations is calculated and treated as an estimate of the audible difference in the two signals.

PESQ attempts to distinguish between impairments that have limited influence over perception and those that have greater effect. It applies lower levels of compensation for those distortions that have less effect on human perception and greater levels of compensation for distortions that have a larger anticipated effect. Ultimately, audible differences are aggregated to generate a single MOS score.

A follow-on recommendation for speech quality analysis, Perceptual Objective Listening Quality Analysis (POLQA), has been approved (ITU-T Recommendation P.863) and initial capabilities are under development. The RT&E Center did not have access to these tools for use during this study.

A.5 MEASURING VIDEO QUALITY

There are a number of tools for measuring video quality.

A.5.1 PSNR

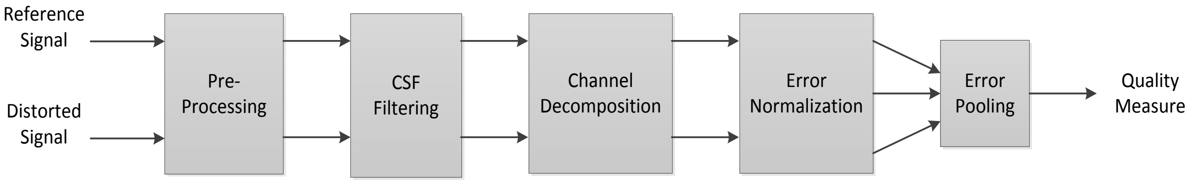
The simplest form of full-reference objective evaluation of image or video quality is to perform a pixel by pixel mathematical comparison of the input and output files, and the two most common metrics for expressing the results of these mathematical files are Mean-Square-Error (MSE) and PSNR. MSE, as the name implies, is computed by taking the mean value of the squares of the deviations between all the pixels in the input and output files. PSNR uses MSE to compute the ratio between the maximum possible signal power and the power of the corrupted part of the signal to estimate the degree of degradation in the image or video.

Unlike other metrics available for expressing image or video quality, both MSE and PSNR are unbounded values, making them harder to use. PSNR is nominally easier to use than MSE, since like other metrics, higher values of PSNR correspond to higher fidelity images or videos. Both

MSE and PSNR provide accurate expressions of network performance, making them especially useful for evaluating technical systems and identifying potential improvements. However, because they evaluate quality in a manner very different from the way human beings interpret images, they are less helpful in monitoring and measuring dynamic changes in quality of experience across a network.

A.5.2 Perceptual Approaches

Metrics such as MSE and PSNR assess pixels within an image individually and view them as independent from one another; but that is not how humans view images. Humans cannot discriminate individual pixels, and they apply a more holistic interpretation of images. As a result, simply measuring the magnitude of the distortion between two images may not reflect the true level of distortion. Equivalent levels of actual distortion can have vastly different effects on human perception, depending on where and how the distortion is distributed. Perceptual approaches to quality assessment attempt to measure quality in a manner more consistent with the way humans view images. Figure A–1 contains a representation of a generic model of a system for performing quality assessment based upon error sensitivity.



Adapted from Wang et al., “Image Quality Assessment.”

Figure A–1: Quality Assessment Framework

Within the process described in Figure A–1, there are five steps in the process of assessing video quality:

1. Pre-Processing: During pre-processing the reference and distorted signals are spatially aligned (just as moments of speech and pauses must be aligned in an audio stream, structures captured within an image must be aligned) and scaled. In addition, a number of transforms, referred to as wavelet transforms, are executed on the input signal in order to reduce dependences within the signal.
2. Contrast Sensitivity Function: Luminance values are used to calculate contrast within an image. In perceptual assessment approaches, computation of contrast values will be aligned with the sensitivity of the human eye to various levels of spatial and temporal difference. This function can be performed prior to channel decomposition or as part of the error normalization process.
3. Channel Decomposition: Images are divided into “channels” based upon spatial and temporal frequency and orientation.
4. Error Normalization: For each channel, the difference between the distorted and reference signal (i.e., the “error”) is calculated and normalized. The goal of normalization is to convert the error into measures of Just Noticeable Difference (JND).

5. Error Pooling: Normalized errors over the spatial extent of the image and across channels are aggregated. The aggregation algorithm may weight some portions of the image more highly than others.

A.5.3 SSIM[3]

SSIM was originally developed as a model at the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin, and a full algorithm was later developed with the Laboratory for Computational Vision at New York University. It was designed to predict the perceived quality of digital television and movies. SSIM is an open source algorithm; it is readily described in published literature. MATLAB versions of SSIM are readily available on the internet.

When viewing images, humans cognitively identify groups of pixels and infer structure. Non-structural distortions, such as changes in luminance and contrast or spatial shifts, have limited effect on the ability of human observers to extract critical information from an image, whereas structural distortions, including blurring, noise, and JPEG blocking cause the loss of the critical structural information humans need to interpret an image.

Unlike PSNR and MSE, SSIM measures groups of pixels using a sliding window that is moved across an image. Distortions are measured locally and aggregated for each image. In addition, SSIM divides video quality based upon three features of the images: luminance, contrast, and structural. Luminance, which is what is measured directly by a camera, is a product of both the illumination on the structures in an image and the reflectivity of those structures. What SSIM attempts to do is to isolate the effects of illumination from the structural information contained within each image of a video.

Figure A–2 provides a top-level view of SSIM processing. In this process, signals x and y represent the images or video streams to be compared. SSIM performs its measurements, using an 8 by 8 pixel (or 11 x 11) sliding window, which moved across an image pixel by pixel, beginning in the upper left-hand corner and ending in the bottom right. Localized windows work better for a number of reasons: key features of the image may not be stationary; distortions may be spatially distributed and human observers do not tend to view an entire image at once but move from localized window to localized window across the image. A circular-symmetric Gaussian weighting function applies in order to smooth each of the localized images.

[3] Wang et al. "Image Quality Assessment."

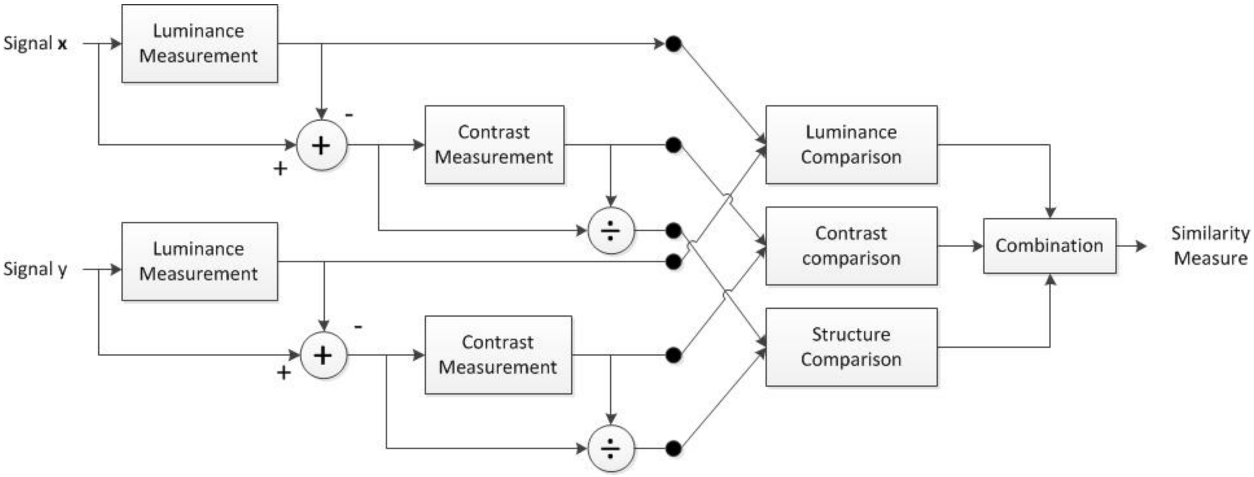


Figure A–2: SSIM Processing[4]

First, the luminance of each signal is computed as the mean intensity of the signal; the luminance values derived from each is used in the luminance comparison. Second, the mean intensity is subtracted from the signal and the resulting standard deviation is used as an estimate of the contrast. The resulting contrast values from each of the two signals are used as inputs to the contrast comparison. Third, each signal is divided by the standard deviation computed during the contrast measurement and input to the structure comparison. The three resulting measures: luminance difference, contrast difference and structure difference are combined to provide an estimate of the similarity of the two signals.

Comparison functions were developed for each of these tests. In developing these tests three criteria were applied:

- Comparisons must by symmetric: S(x,y) = S(y,x)
- Comparisons must be bounded: -1 ≤ S(x,y) ≤ 1
- Comparisons must have a unique maximum: S(x,y) may only equal 1, if x = y.

The luminance and contrast comparison functions take the same form:

$$f(x,y) = \frac{2xy + C}{x^2 + y^2 + C}$$

Constants are added to the numerator and denominator to avoid division by zero. The structure comparison is defined as follows:

$$s(x,y) = \frac{\sigma_{xy} + C}{\sigma_x \sigma_y + C}$$

[4] Wang et al. “Image Quality Assessment.”

Finally, the three comparison functions are combined:

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma}$$

The three constants, α, β and γ allow different weights to be applied to the luminance, contrast and structure functions. The specific form of this equation is captured in the following:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

A.6 APPLICATION

Software used to compute PSNR and SSIM during the execution of this study were downloaded from OpenCV (Open Source Computer Vision Library). This open source computer vision and machine learning software library includes a comprehensive set of over 2500 optimized computer vision and machine learning algorithms. The software is freely available under the terms and conditions of a Modified Berkley Software Distribution (BSD) license. OpenCV has a user community of more than 47 thousand and an estimated number of downloads exceeding 14 million.[5]

OpenCV is designed to run on a variety of platforms and operating systems.[6] The VTC project selected two unique platforms for purposes of comparison and validation. Either one of these two will be sufficient in order to perform the calculations outlined in the following sections.

Table A–1: Specifications for Two Different Platforms for Computing Objective Video Quality Measurements

Platform	MacBook Pro	Dell Latitude E6530
OS	OS X Version 10.11.6	Windows 7 Enterprise
System Type	64-bit OS	64-bit OS
Processor	Intel Core i7 @ 2.8 GHz	Intel Core i7-3740QM @ 2.7 GHz
Memory	16 GB 1600 MHz DDR3	8 GB
Display Adapter	AMD Radeon R9 M370X 2048 MB	NVIDIA NVS 5200M 1024 MB
Compiler	Apple LLVM 9.0.0	LLVM 5.0.0

[5] OpenCV (web page).

[6] Please refer to the OpenCV Wiki for detailed information and instructions on the proper installation and configuration of the software library, https://github.com/opencv/opencv/wiki.

A.6.1 Obtaining the Software/Establishing an Operational System

Downloading the software and establishing an operating system to execute the software can be achieved as follows:

1. Install a compiler. Although there is not a specific requirement for this choice, the LLVM compiler was selected because it could support both the Windows and Operating System (OS) X platforms. Detailed information and instructions can be found on the website[7]
2. Install the OpenCV library. The software (version 3.2) can be downloaded from the website www.opencv.org. The file obtained should match the specifications in Table A–2 with emphasis on the checksum calculation. Verifying the checksum of the downloaded file matches the value in the table will confirm its integrity.

Table A–2: Specifications for Software File Downloaded from OpenCV Website

OS	OS X	Windows
File URL	https://github.com/opencv/opencv/archive/3.2.0.zip	https://github.com/opencv/opencv/releases/download/3.2.0/opencv-3.2.0-vc14.exe
Checksum (MD5)	bfc6a261eb069b709bcfe7e363ef5899	7631e708a9ae036569e400ba43886861
Size (MB)	78 (Zipped), 145 (Extracted)	118

A.6.2 OpenCV Overview

The OpenCV distribution includes a set of tutorial programs designed to introduce the new user to the framework, and the interaction data structures and objects and how they interact in order to provide meaningful results. The set of tutorials included with the distribution is comprehensive and highlights some state-of-the-art image and video processing techniques.

The OpenCV BSD license allows the developer to duplicate and modify the source file video-input-psnr-ssim.cpp in order to extend its functionality to provide meaningful results. Meaningful results for the VTC project are the calculation and collection of SSIM, and PSNR values for a series of reference and test video clips.

A.6.2.1 Tailoring the Program

A number of small revisions had to be made to the OpenCV software to tailor it to the needs of this project. The most significant changes made affected the way frames were aligned between the original reference file and the distorted output file. There were also smaller changes to modify the format of output files.

[7] LLVM Compiler Infrastructure (website), https://llvm.org.

A.6.2.2 Frame Count

Frame counts are correct if, and only if, there is a one-to-one relationship between the set of reference frames and the set of test frames (i.e., no video frames are lost during transmission). When there is a one-to-one relationship between the frames in the reference and distorted videos two VideoCapture objects can read image frames from the videos using the same frame counter resulting in efficient iteration and computation. Unfortunately, this is usually not the case.

The test bed is designed to transmit videos under varying network loads and regularly conducts test scenarios that include high loading and packet loss conditions. These conditions will ultimately lead to a reduction of the number of video frames that are captured and recorded in the test video versus the number of frames that were originally transmitted. A simplified illustration of this phenomenon is shown in Figure A–3.

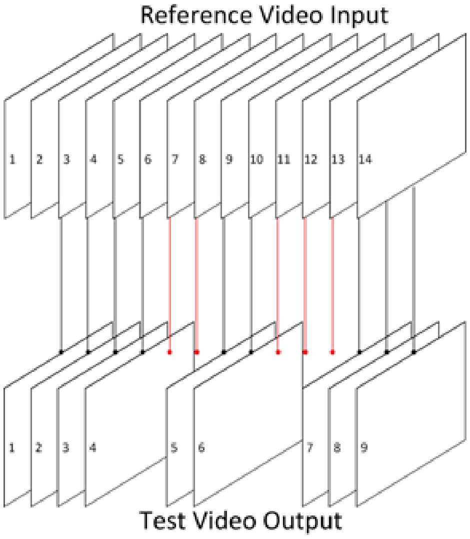


Figure A–3: Frame Loss Illustrated

If there is no compensation, the loss of frames creates a situation in which the measurement tool is unable to temporally align the frames in the distorted output video clip to the corresponding frames in the reference video. To compensate for lost frames, the RT&E Center study team modified the measurement tool to compare each frame in the output video to a number of adjacent frames in the reference video. The frame in the reference video with the highest level of correlation was assumed to be the corresponding frame in the distorted video. Overall, each frame in the distorted video was compared to 21 adjacent frames. The processing is illustrated in Figure A–4. This added processing was computationally intensive. Table A–3 summarizes the resulting run times for each clip. Each series contains 24 sub clips and each sub clip contains approximately 610 frames.

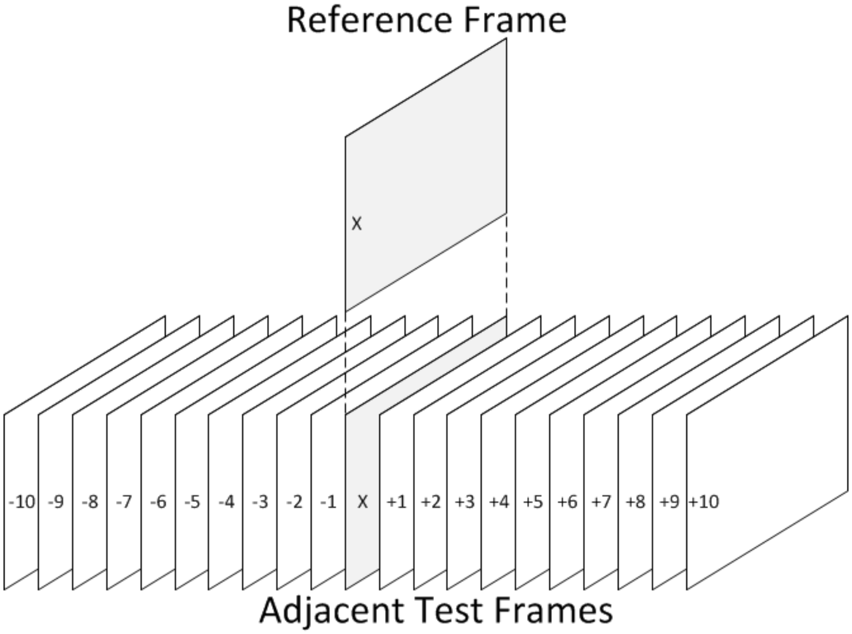


Figure A–4: The Program Performs 21 Calculations for SSIM and 21 Calculations for PSNR for Every Reference Frame

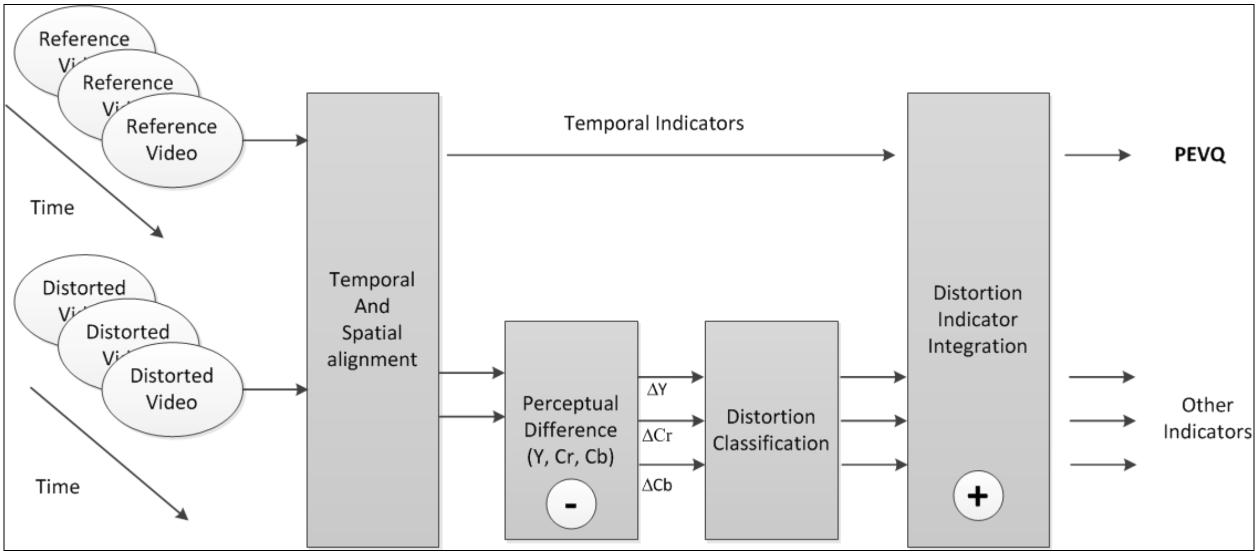
Table A–3: Run Times for Computation

Platform	Clip Runtime (min.)	Series Runtime (hr.)
MacBook Pro	40	16
Dell Latitude E6530	55	22
Dell Latitude E6530, NVIDIA GPU	20	8

A.7 PEVQ

The final tool to be evaluated as part of this study is PEVQ. PEVQ is a full-reference measurement tool that performs objective assessments of video quality and converts the results into MOS. PEVQ has been benchmarked by the Video Quality Experts Group (VQEG).

PEVQ is a full-reference perceptual measurement algorithm that performs pixel analysis of corresponding frames within two videos to generate an assessment of the perceptual quality of the output video. Degradations and artifacts resulting from coding of the video for network transmission are assessed using models of human visual perception. Results of these analyses are converted into a MOS, which have been benchmarked against subjective assessments from human subject testing. In addition, PEVQ records other indicators including PSNR and lip-sync delays. Figure A–5 contains a top-level representation of the PEVQ process.



Adapted from Opticom, http://www.pevq.com/pevq.html.

Figure A–5: PEVQ Algorithm Description

The basic process can be described in the following steps:

- A temporal and spatial alignment is performed to ensure that the algorithm is comparing corresponding frames in the input and output files.
- Perceptual differences between the input and output files are determined. Unlike PSNR measurements that compute pixel differences and aggregates them, PEVQ uses a model of the human visual system to identify only those differences that are detectable by the human eye.
- A temporal information indicator is used to provide the algorithm with an indication of whether the action in the frames being measured is quick motion or relatively static. Human perception varies depending on the pace of the action being observed.
- Previously calculated indicators and distortions are classified.
- The results of the previous steps are aggregated to create a MOS value between 1 and 5.

When PEVQ is used to assess the quality of impaired files that have been transmitted through a network, the video quality estimation includes impacts from both packet level impairments (loss and jitter), and distortions caused by the coding processes (blockiness, jerkiness, and blur).

PEVQ’s performance was assessed by an independent third party during standardization benchmarks by the VQEG. It became part of the ITU-T Recommendation J.247 in 2008.[8]

PEXQ™ is a stand-alone product in a family of perceptual quality assessment analyzers developed by Opticom.[9] PEXQ integrates multiple quality assessment algorithms into a single

[8] ITU-T Rec. J.247.

[9] Opticom is located in Erlangen, Northern-Bavaria, Germany. They have been providing standards based quality measurement tools since 199. Opticom’s solutions have been used by various network operators, original equipment

package. This includes the capability of assessing voice, audio/visual and data payload analysis based on human perception. PEXQ specifically uses the PEVQ Analyzer to measure perceived video quality.

A.7.1 PEXQ Usage Overview

PEXQ combines several different perceptual quality assessment libraries for video and audio. It also provides a graphical user interface (GUI) that interfaces with the assessment libraries permitting the calculation and aggregation of quality scores. PEVQ operates through a command line interface (CLI) language instead of a GUI. Command line interface language is similar to scripts and provides direct control over parameters and algorithms.

Running PEVQ using CLI may have an entirely different syntax compared to PEXQ. Refer to the documentation that comes with the software for details on the command line interface for PEVQ.

A.7.1.1 System Requirements

Table A–4 contains the minimum requirements and the specifications of the computer used to run PEXQ.

Table A–4: Requirements and Specification of PC Hardware used for PEXQ

	Minimum PC Requirements	Dell OptiPlex 9020
OS	Windows XP, Windows Vista, Windows 7	Windows 7 Enterprise
Processor	PC with Pentium III 500 MHz or Better	Intel Core i7-4770 CPU @ 3.40 GHz 4 Core(s), 8 Logical Process(s)
Memory	1 GB of RAM	16 GB
Hard Disk Space	1 GM of disk space for the PEXQ application and related files	320 GB with 40 GB of free disk space
.NET Framework	Microsoft .NET framework version 2.0 with service pack 2	Microsoft .NET framework version 4.6
Hardware Interfaces	USB/Parallel Port (if hardware key/dongle required)	USB 3.0

manufacturers and in R&D. Opticom.de. (2017) OPTICOM perceptual voice audio quality test products OEM technology, accessed 13 Nov. 2017, http://www.opticom.de/company/customers-voice-market.php

A.7.2 Input File Requirement

Input requirements include: Raw YUV, Audio Video Interleave (AVI) files with RGB24, YUV444, YUV422, or YUV420 data,[10] frame rates from 2.5 to 60 frames per second; video test segments length from 6 to 20 seconds in duration. Ten-second segments were chosen to obtain ITU-T P.910[11] compliant MOS scores.[12]

A.8 CALCULATING MOS WITH PEVQ/PEXQ

Video quality assessments were performed using PEXQ and its PEVQ library. PEXQ is capable of running PEVQ analysis as batch jobs using CLI. This is more productive when more than one file pair requires assessment. The instructions below will highlight the workflow that was used to assess multiple reference/impaired video pairs by using CLI.

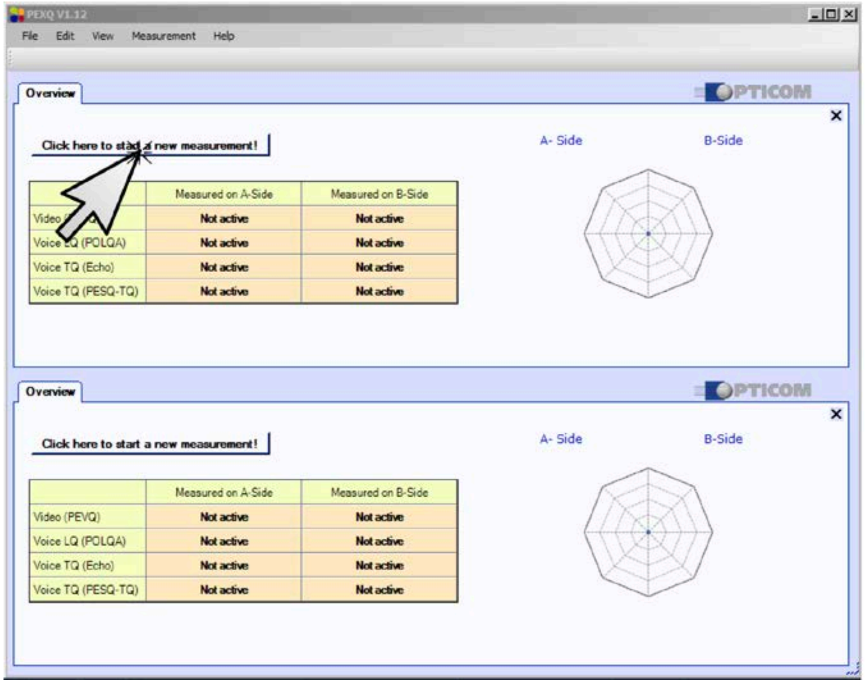
The first step is to locate the PEXQ application on the windows machine. After locating the program, launch it by doubling clicking on the icon.

Upon program startup, a window will appear with options to start the measurement. Press on the “Click here to start a new measurement” button to begin selecting the reference and impaired input files. Ignore the second option, it is a duplicate window.

[10] Audio Video Interleave is an audio/video file container that can hold uncompressed RAW YUV format. YUV is a color space pixel format that contains bit map images in three components; luminance (Y), and color (U,V). Red, Green and Blue (RGB) is another color space designed for computer displays where each pixel contains Red, Green and Blue components to render color images.

[11] ITU-T Rec. P.910.

[12] Additional information on PEVQ can be obtained from http://www.pevq.com/pevq.html.

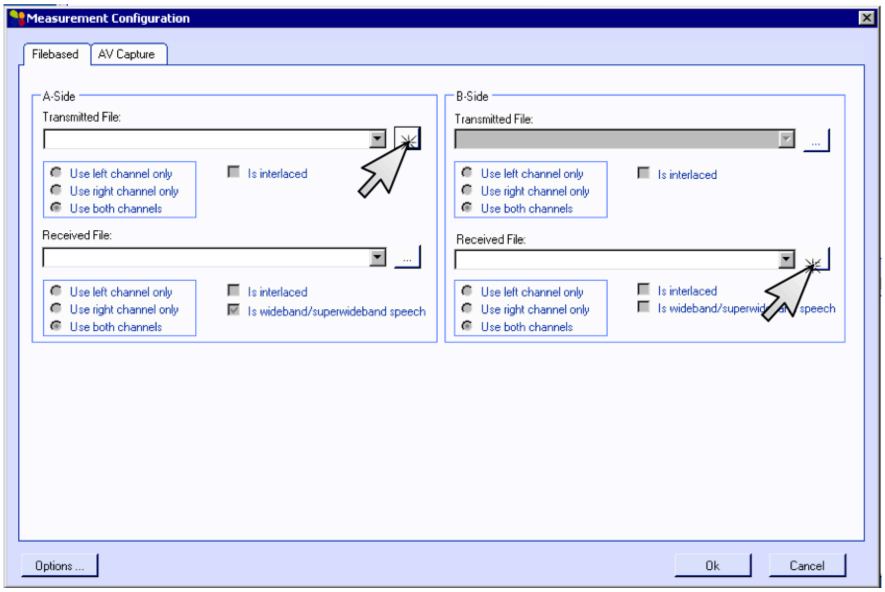


	Measured on A-Side	Measured on B-Side
Video (PEVQ)	Not active	Not active
Voice LQ (POLQA)	Not active	Not active
Voice TQ (Echo)	Not active	Not active
Voice TQ (PESQ-TQ)	Not active	Not active

	Measured on A-Side	Measured on B-Side
Video (PEVQ)	Not active	Not active
Voice LQ (POLQA)	Not active	Not active
Voice TQ (Echo)	Not active	Not active
Voice TQ (PESQ-TQ)	Not active	Not active

Figure A–6: PEXQ Screenshot to Start New Measurement

After the new measurement button is pressed, a window will appear with options to select input files for A-side transmitted/received file and B-side transmitted/received file. A-side transmitted is for the reference file and the B-side received is for the Impaired file.



A-Side

Transmitted File:

○ Use left channel only
○ Use right channel only
● Use both channels

□ Is interlaced

Received File:

○ Use left channel only
○ Use right channel only
● Use both channels

□ Is interlaced
☑ Is wideband/superwideband speech

B-Side

Transmitted File:

○ Use left channel only
○ Use right channel only
● Use both channels

□ Is interlaced

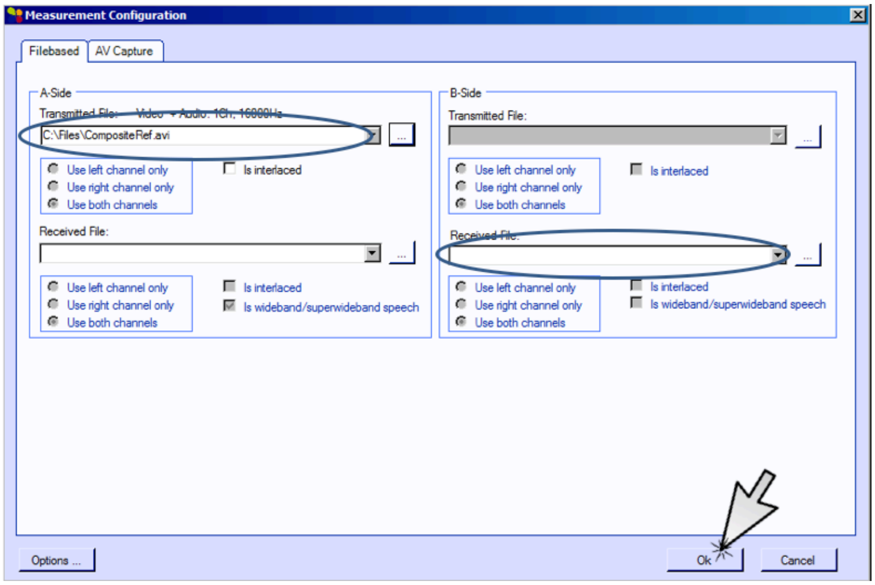
Received File:

○ Use left channel only
○ Use right channel only
● Use both channels

□ Is interlaced
□ Is wideband/superwid speech

Figure A–7: PEXQ Screenshot to Select Input Files

Select the desired reference file and impaired file by clicking on the button next to the box for the file name on A-side Transmitted File and B-side Received file as indicated above. Once the selections are made, the Measurement configuration window will display the selected file as shown below. A-side video should be the reference video and B-side input should be the impaired video.



Measurement Configuration

Filebased | AV Capture

A-Side

Transmitted File: Video + Audio: 1Ch, 16000Hz

C:\Files\CompositeRef.avi

○ Use left channel only
○ Use right channel only
● Use both channels

□ Is interlaced

Received File:

○ Use left channel only
○ Use right channel only
● Use both channels

□ Is interlaced
☑ Is wideband/superwideband speech

B-Side

Transmitted File:

○ Use left channel only
○ Use right channel only
● Use both channels

□ Is interlaced

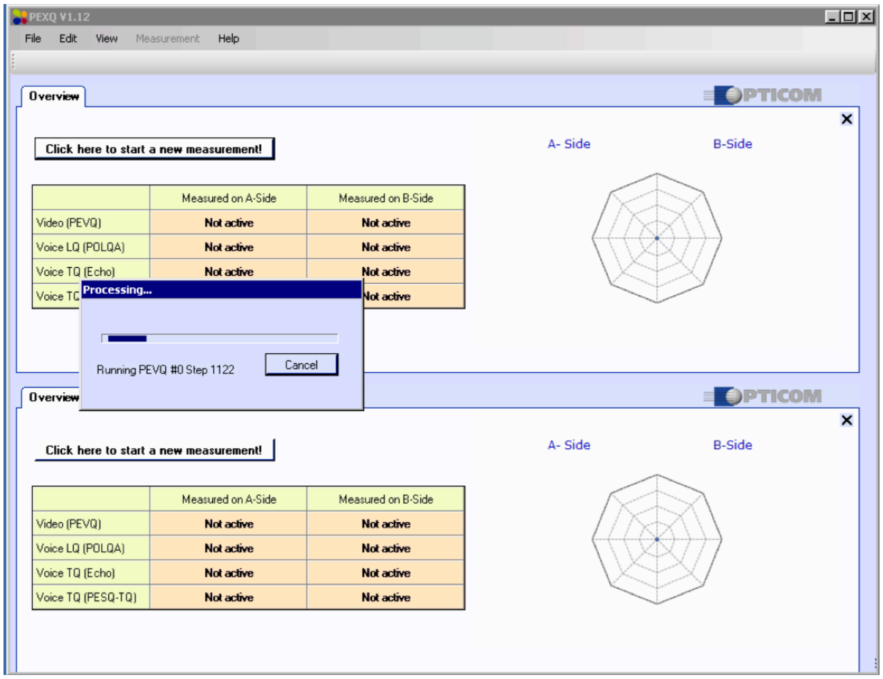
Received File:

○ Use left channel only
○ Use right channel only
● Use both channels

□ Is interlaced
□ Is wideband/superwideband speech

Figure A–8: PEXQ Screenshot to Display Selected Files

After both inputs are selected, press "OK" to start the assessment calculation.



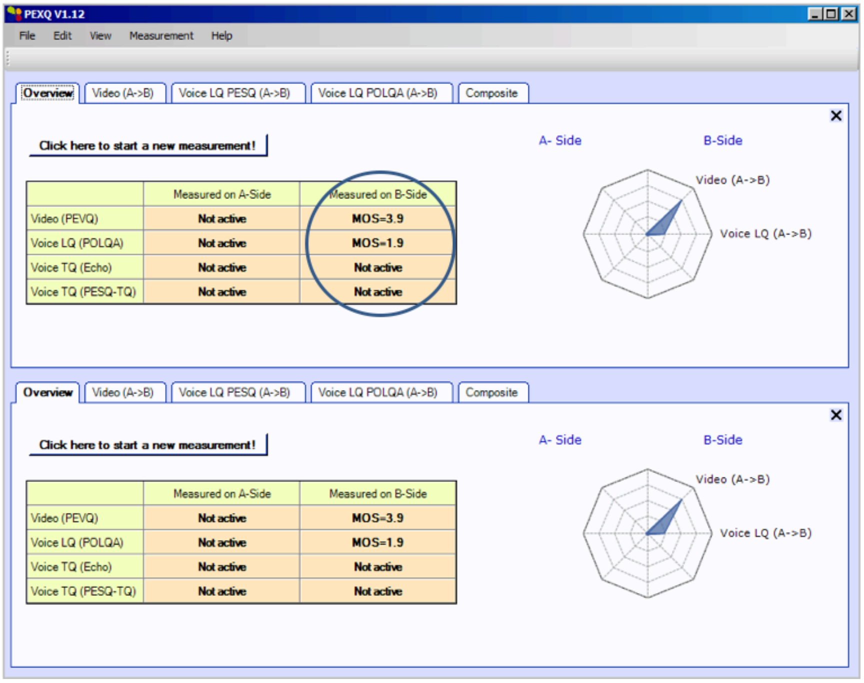
	Measured on A-Side	Measured on B-Side
Video (PEVQ)	Not active	Not active
Voice LQ (POLQA)	Not active	Not active
Voice TQ (Echo)	Not active	Not active
Voice TQ (PESQ-TQ)	Not active	Not active

Processing...

Running PEVQ #0 Step 1122

Figure A–9: PEXQ Screenshot to Start the Assessment Calculation

The PEVQ assessment will now run for some time and when completed, the results will be shown on the dashboard as illustrated below.



	Measured on A-Side	Measured on B-Side
Video (PEVQ)	Not active	MOS=3.9
Voice LQ (POLQA)	Not active	MOS=1.9
Voice TQ (Echo)	Not active	Not active
Voice TQ (PESQ-TQ)	Not active	Not active

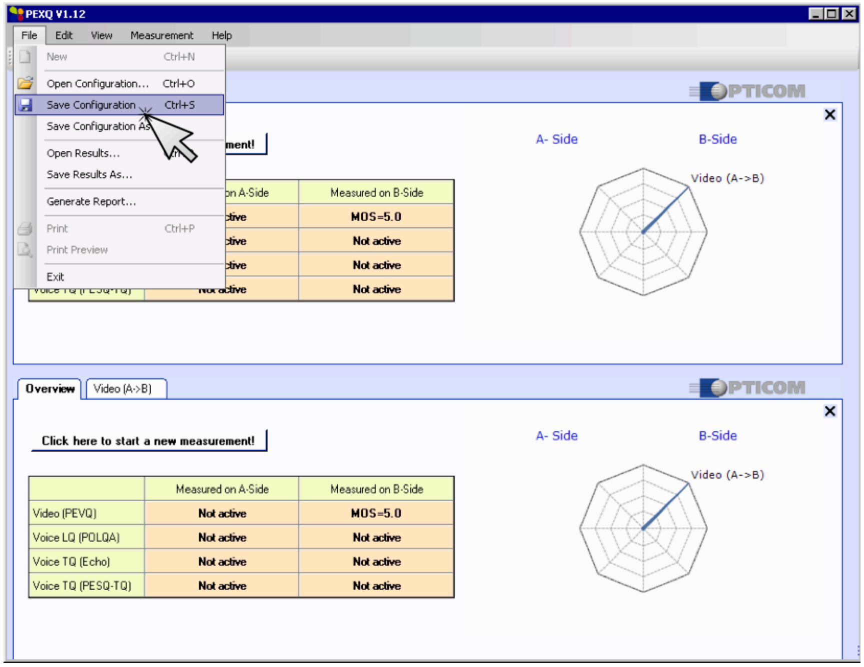
	Measured on A-Side	Measured on B-Side
Video (PEVQ)	Not active	MOS=3.9
Voice LQ (POLQA)	Not active	MOS=1.9
Voice TQ (Echo)	Not active	Not active
Voice TQ (PESQ-TQ)	Not active	Not active

Figure A–10: PEXQ Screenshot to Display the PEVQ Assessment

The workflow so far is one input pair. PEVQ Video quality assessment using PEXQ is a manual process and must be completed for each input pair. This is especially time consuming for assessments involving multiple file pairs. The analysis could be performed using CLI scripts to continuously run PEVQ in between video pairs without manual intervention. Since this study focused on assessing a large number of input pairs, CLI was used to run the video quality assessment instead of running the assessment one at a time using PEXQ.

In order to perform CLI batch jobs for PEVQ, a special configuration file that defines the A-side, B-side parameters, type of analysis and other parameters is requested. PEXQ can be used to generate that configuration file. The steps below describe that process.

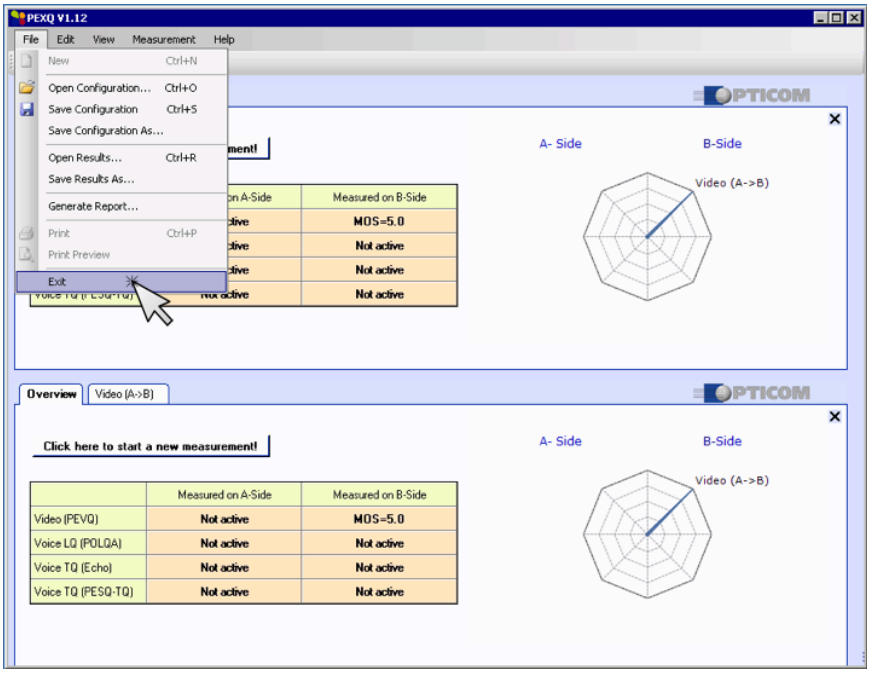
Once the PEVQ analysis completed with results on the dashboard, as above, go to the File menu and select “Save Configuration” from the drop-down menu. Name and save the configuration file to a location that can be easily remembered and accessed.



	Measured on A-Side	Measured on B-Side
Video (PEVQ)	Not active	MOS=5.0
Voice LQ (POLQA)	Not active	Not active
Voice TQ (Echo)	Not active	Not active
Voice TQ (PESQ-TQ)	Not active	Not active

Figure A–11: PEXQ Screenshot to Save Configuration File

After the file is saved, exit PEXQ by going to the File menu and then select “Exit” from the drop-down menu.



	Measured on A-Side	Measured on B-Side
Video (PEVQ)	Not active	MOS=5.0
Voice LQ (POLQA)	Not active	Not active
Voice TQ (Echo)	Not active	Not active
Voice TQ (PESQ-TQ)	Not active	Not active

Figure A–12: PEXQ Screenshot to Exit the Program

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

With the creation of the configuration file, CLI can then be used to run PEVQ assessment in batches through the command prompt window.

A.8.1 Tailoring the Program

PEXQ can be run using the script below in the window's command prompt.

```
PEXQ -Config c:\configfile.xml -FileTxA c:\reference.avi -
FileRxB c:\impaired.avi -Out c:\assessmet.txt
```

Where *c:\configfile.xml*, *c:\reference.avi* and *c:\impaired.avi* are the names of the files and the directories where the configuration file, the reference file and impaired files are located. The assessment results are then written to a text file named *assessment.txt* located in the c: directory. The script above is for one video pair. There is no requirement for file name or their location, as long as the directory paths indicated in the script exactly match the actual name and location.

For multiple video file pair assessment, the basic PEXQ script above can used to create a batch file using a text editor by making appropriate file name and location changes as shown below.

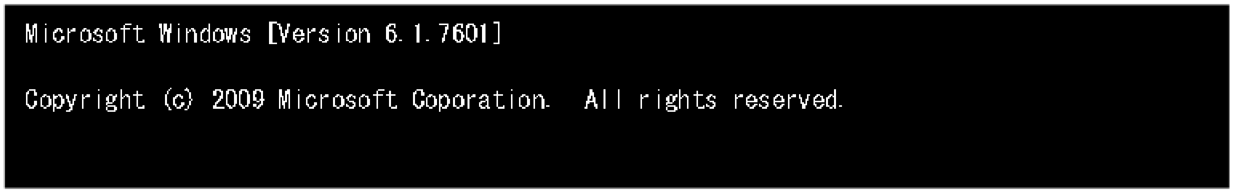
The script with multiple file pairs can then be saved as a .bat file to be executed in the command prompt window. For example, after a file named inputpair.bat is saved to the root directory c: the batch job can be initiated in the command prompt window, as illustrated below, to run analysis on all the video pairs without manual intervention.

```
PEXQ -Config c:\configfile.xml -FileTxA c:\reference.avi -
FileRxB c:\impaired1.avi -Out c:\assessmet1.txt

PEXQ -Config c:\configfile.xml -FileTxA c:\reference.avi -
FileRxB c:\impaired2.avi -Out c:\assessmet2.txt

PEXQ -Config c:\configfile.xml -FileTxA c:\reference.avi -
FileRxB c:\impaired3.avi -Out c:\assessmet3.txt

PEXQ -Config c:\configfile.xml -FileTxA c:\reference.avi -
FileRxB c:\impaired4.avi -Out c:\assessmet4.txt
```



```
Microsoft Windows [Version 6.1.7601]

Copyright (c) 2009 Microsoft Coporation.  All rights reserved.
```

A.8.2 Output

Once the PEVQ calculations are completed through the PEXQ CLI, the results are saved into an output file located and named as scripted in the CLI above. From the output file, various measurements can be obtained, including the following:

- PEVQ MOS: Ranging from 1 (bad) to 5 (excellent) is based on multitude of perceptually motivated parameters and is calculated in accordance with ITU-T Recommendation J.247[13].
- Distortion Indicator: Detailed analysis of perceptual level of distortion in the luminance, chrominance and temporal domain.
- Delay: The delay of each frame compared to the reference signal.
- Brightness: The brightness of the reference and degraded signal.
- Contrast: The contrast of the reference and degraded signal.
- PSNR: A coarser analysis of distortion in the Y, Cb and Cr components.
- Jerkiness: Smoothness of the video playback.
- Blur: Reduced sharpness.
- Blockiness: Indication of low bit rate coding.
- Frame Skips and Freezes: Temporal artifacts in video transmission.

A.9 GETTING THE SOFTWARE/ESTABLISHING AN OPERATIONAL SYSTEM

1. The PEXQ suite is no longer supported however PEVQ is still commercially available through Opticom or their distributor. More information regarding the purchase and licensing of the PEVQ software can be obtained by contacting Opticom (http://www.opticom.de/).
2. PEXQ was installed and used to perform video quality assessment. Since PEXQ is no longer supported, detailed instructions on PEXQ software installation will not be relevant, as a result those steps will not be described here. Additionally, the software installation process may be different for PEVQ. It is recommended that the end user follow PEVQ installation instructions in accordance with the installation/user guide after purchase or contact their technical support department.

[13] ITU-T Rec. J.247.

APPENDIX B. VIDEO SELECTION AND PREPARATION

B.1 INTRODUCTION

The RT&E Center's testing was intended to investigate quality issues resulting from network transmission and bandwidth, with particular interest in the ability of VTC to allow users to observe aspects of demeanor such as nonverbal cues, facial expressions and body language which were noted as critical to determining whether a VTC hearing is experienced similarly to an in-person hearing

Demeanor may include the subjects' appearance, behavior, and tone of voice. A survey of trial judges found that credibility was most often based upon "evasiveness, defensiveness, and rationalization" indicated by changes in the witness's behavior.[1] In one study, a judge stated that the ability to observe a participant's demeanor and emotions was a deciding factor when determining whether VTC was an acceptable alternative to an in-person hearing.[2] A case brief included the judge's conclusion supporting use of a VTC hearing, stated that the VTC video quality was "flawless" and that "any hesitation, discomfort, arrogance, or defiance would have been easily discerned."[3]

In contrast, Wellborn stated, "According to the empirical evidence, ordinary people cannot make effective use of demeanor in deciding whether to believe a witness. On the contrary, there is some evidence that the observation of demeanor diminishes rather than enhances the accuracy of credibility judgments."[4]

Nonetheless there is a long case history of judges relying on observation of demeanor when determining credibility.[5] Therefore, it is critical to ensure that the VTC system does not inhibit a judge's ability to observe demeanor.

B.2 OBJECTIVE

The purpose of creating a reference video file and impaired file was to create a set of known video files in order to evaluate objective and subjective video quality measurement techniques in support of the objectives of this research. The workflow was not intended to directly assess codec performance nor was the emulated network designed to replicate an actual VTC network topology.

[1] Timony, "Demeanor Credibility."

[2] Davis et al. "Research on Videoconferencing."

[3] U.S. Court of Appeals, Case No. 15–1349, Document No. 1613347, filed 05/16/2016. 70.

[4] Wellborn, Olin G. III, "Demeanor."

[5] Timony, "Demeanor Credibility," 913.

B.3 MATERIALS AND METHODS

In this test configuration, video is inserted into the Cisco codec and transmitted across an emulated packet data network to a second Cisco codec. Video quality can be degraded by introducing packet losses or variance in packet delays into the emulated network. The test configuration enables users to define a level of packet loss (in terms of percentage of packets lost) or variance in the delay time for received packets (a common cause of jitter) and observe the resulting video distortion. Within the lab, observers can simultaneously observe the input and impaired output streams on a pair of side-by-side video monitors.

A reference video is recorded at the input codec and the impaired output video stream is recorded as an output from the receiving codec, and both streams are fed into the quality analysis tool.

B.3.1 Video Selection Process

For the purpose of this study, the original video as purchased is referred to as the reference video. The reference videos were selected to capture a variety of situations typical of VTC streams from actual hearings.

The Center for Legal & Court Technology's (CLCT) report on best practices for use of VTC in hearings notes that "videoconferencing that is well designed from the outset and that follows the best recommendations … should in no way prevent judges from making credibility decisions over videoconferencing…" It goes on to suggest that a judge's perception of credibility may be increased when the VTC includes the ability to zoom in to see details otherwise undetectable from the bench.[6] However, the assumption was that not all VTC-enabled courtrooms would be fully compliant with best practices.

Because the study was interested in the effects of video quality that might impact the outcome of a hearing, it was important to include a range of scenes representative of situations that could reasonably be expected in a courtroom setting, including settings that do not follow best practices for room design and lighting. Inclusion of test videos with poor initial quality also allowed for study of whether a specific level of degradation caused an equivalent drop in quality when applied to videos filmed under both good and poor conditions.

B.3.1.1 Quantitative Criteria

Aspects of the test bed influenced the selection of the reference videos. For example, the evaluation of MOS scores required 10-second test clips. Therefore, the initial videos needed to be long enough to trim to 10 seconds.

[6] CLCT, "Best Practices for Using Video Teleconferencing," 14.

There was initial concern that 10 seconds would not allow human participants enough time to make subjective assessments of the video clips, but Willis and Todorov[7] found that people formed judgments about factors such as trustworthiness and competence after only a 100-ms exposure to unfamiliar faces, and that this exposure time was not only sufficient for participants to form an impression. Increased exposure time did not significantly change those initial opinions.

Reference videos were selected to mimic expected output from courtroom VTC video. Based on the literature review of the equipment and video output, reference videos were required to meet the following quantitative criteria:

- Be in full color
- Have duration of more than 10 seconds
- Be in video format MP4
- Have video resolution of 1920 x1080 or 1280 x 720 high definition (HD)
- Have a frame rate of approximately 30 frames per second

B.3.1.2 Qualitative Criteria

Because the same set of videos would be used for both objective and subjective testing, the study team identified video clips that would test conditions that can affect viewer's perception of demeanor and might reasonably be found in a courtroom. Conditions include variation in contrast between subject and the background; variety of skin tone, gender and age; and apparent mood or expression exhibited by the subject. Additional elements that might affect the VTC system's ability to digitally capture and display information were also considered, including initial white balance, angle of the lighting on the subject, distance or angle between camera and subject, and reflective surfaces.

Contrast between Skin Tone and Scene Background

- Apparent light, medium, or dark complexion
- Apparent light, medium, or dark background
- Single color or variegated background

Room Setup

- Using best practices for lighting level and color balance
- Lighting level too high (subject is over exposed)
- Lighting level too low (subject is under exposed)
- Color balance results in unnatural skin tone
- Lighting position causes unnatural shadows on the face

[7] Willis, J., and A. Tudorov, "First Impressions."

Motion

- Subject is relatively still against a stationary background
- Subject displays hand / arm gestures
- Subject is relatively still against a moving background
- Subject is moving against a stationary background

Gender of Video Subject as Compared to Gender of Study Participant

- Male evaluating male
- Male evaluating female
- Female evaluating male
- Female evaluating female

Role Displayed by Subject

- Decision-maker / Authority (judge, jury, police, lawyer)
- Defendant / Accused
- Victim
- Witness

Apparent Emotions Displayed by Subject

- Neutral
- Confident
- Angry / Frustrated / Hostile
- Fearful
- Sad / Depressed

Facial Features

- Glasses
- Facial hair
- Facial piercings
- Wrinkles
- Tooth loss
- Deep set eyes
- Epicanthic fold

The study duration did not allow for comprehensive investigation of all the identified qualitative test conditions. Therefore, the study team elected to focus on test videos that included conditions which literature review suggested were likely to be experienced in a courtroom setting and could have significant impact on video quality.

The contrast between the room background and the subject of the video can impact the video quality for the end user.[8] When humans look at a scene, they tend to scan individual areas throughout the scene. Human eyes adjust as they change focus, allowing them to interpret detail in both bright and dark areas of a scene. Human brains then combine the information from a series of focal areas into a detailed image of the entire scene.

A VTC video consists of a series of still images, each of which must capture the entire scene. Both the camera and the codec must average the light throughout the scene to capture and transmit the best overall image. This means that they may lose details in areas that are much brighter or much darker than the overall image. This can cause the facial features to be lost when subjects with dark skin are filmed against a bright background. Conversely, reflections appear as bright areas without detail. Based upon the importance of contrast in identifying facial features and the known challenge that auto-contrast adjustment presents for VTC equipment, contrast was identified as the primary test condition for study.

B.3.1.3 Source for Reference Videos

Given the need to evaluate a variety of conditions, the option of obtaining actual courtroom video was eliminated due to difficulty locating multiple jurisdictions willing to share actual courtroom video, particularly video not filmed using best practices. Purchasing stock video was substantially less expensive than producing sample videos. The study team therefore elected to purchase stock footage representative of courtroom situations.

Candidate videos were identified on Shutterstock.com. The initial selection criteria were concerned with finding videos with settings and situations similar to a courtroom, thus videos that included a person seated or standing, addressing the camera, with limited movement, and no background movement.

For the purpose of this study, the team elected to focus primarily on conditions that could be addressed easily through application of best practices for system and room setup. Contrast was the primary test condition, with additional consideration for distractions such as harsh lighting, reflections or shadows, and poor white balance.

Processing time limited the number of videos, which could be evaluated during the study period. This was balanced by the need to include enough variety among the reference videos to avoid test participants becoming overly accustomed to the subjects and thus less sensitive to the perceived video quality. Based on the matrix shown in Table B–1, the team determined that a set of nine videos would be sufficient to capture the range of subject-to-background contrast, as well as offering the opportunity to study some of the lighting variations that could conceivably occur in a courtroom setting.

[8] CLCT, “Best Practices for Using Video Teleconferencing,” 39–40.

Table B–1: Apparent Contrast between Skin Tone and Background

	Light	Medium	Dark
Light	Low contrast	Moderate contrast	High contrast
Medium	Moderate contrast	Low contrast	Moderate contrast
Dark	High contrast	Moderate contrast	Low contrast

B.3.1.4 Subject Matter

Ultimately, the nine reference video clips shown in Figure B–1 were selected to include varied amounts of contrast between the primary subject’s skin and the background while also including variation in gender, age, facial orientation of subject, and whether the subject was wearing eyeglasses.

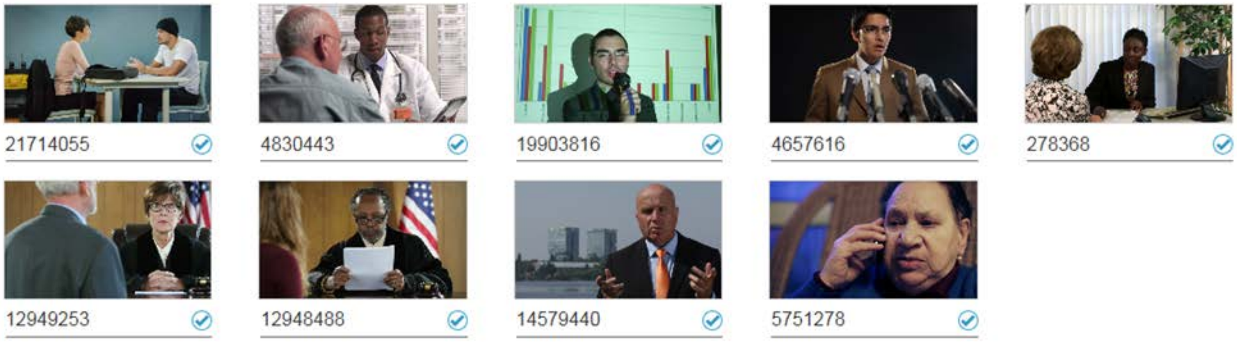
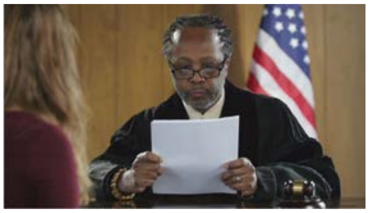
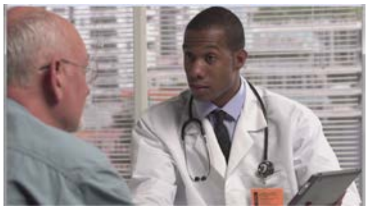
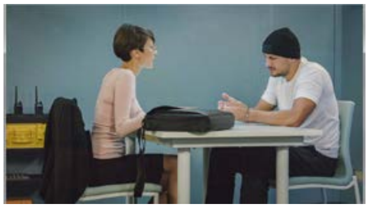
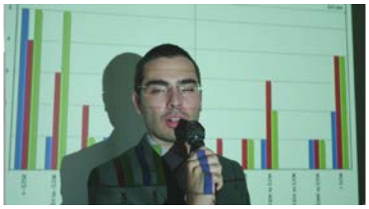
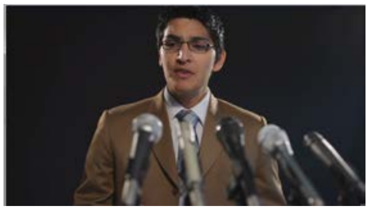


Figure B–1: Videos Selected to Represent a Variety of Test Conditions

During initial processing, it was determined that several reference videos were in an unexpected aspect ratio that would require an additional conversion step. These videos were disqualified rather than introduce an additional variable to the process. Due to time constraints, the study was conducted using only five reference videos, which captured the test conditions described in Table B–2.

Table B–2: Screenshots, Descriptions and Test Conditions Associated with Each Reference Video

Clip ID		Primary Test Condition	Secondary Test Condition(s)
A		Moderate contrast: dark skin tone / moderate background	reflections facial hair
F		High contrast: dark skin tone / light background	oblique facial orientation striated background
G		Moderate contrast: light skin tone / medium background	increased distance between camera and subject body language
H		Low contrast: medium skin tone / medium background	poor white balance improper lighting reflections
I		Moderate contrast: medium skin tone / dark background	improper lighting foreground objects

B.3.2 Video Impairment Process

JHU/APL's Advanced Networking Technologies Lab Hardware in the Loop Test Bed (ANT-HIL) consists of both physical and virtualized networking hardware and functionality. It is a Linux-based environment hosted on a Dell PowerEdge R720 server running VMware ESXi and allows the creation of various emulated packet data networks with customizable topologies and characteristics. Additionally, the ANT-HIL server is connected to a Juniper EX4200 switch to allow connections between the virtualized networks, physical hardware, and other networks. Cisco Tandberg C60 VTC endpoints[9] are connected to the Juniper EX4200 switch through a Cisco Catalysts 3750G switch located in the ANT-HIL lab, and packets sent between the two

[9] https://communities.cisco.com/docs/DOC-47410.

VTC endpoints are routed through the emulated network. Table B–2 represents the network topology used for the emulated environment.

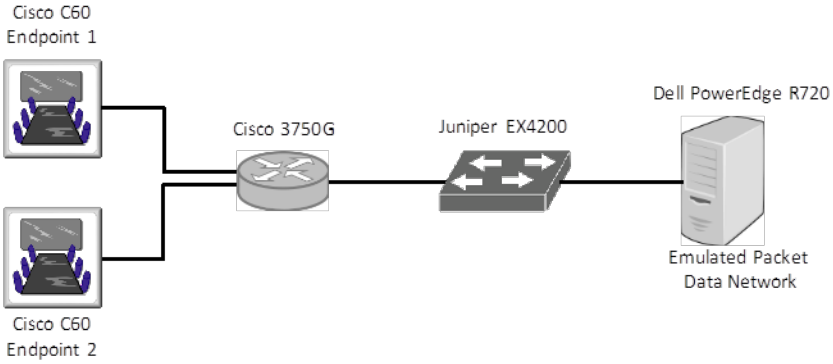


Figure B–2: Emulation Network Topology

The emulated packet data network uses the network emulation (NetEM) kernel module to emulate wide area network (WAN) link characteristics. Network impairment characteristics, such as delay, packet loss, and other variables can be added to outgoing packets on any interface in the emulated network.[10] This allows hardware under test to be physically collocated yet still appear to be communicating over a WAN or other realistic network topology. Controlled testing in this environment provides insight into the effects of the network characteristics on end-to-end application performance. A summary of the video manipulation process is included here.

Two video distortion types that can occur when video is transmitted over a computer network and that are known to negatively impact quality were selected for investigation.
Corruption of synchronization signals or electromagnetic interference during video transmission causes video jitter, exhibited by randomly displaced horizontal lines in the video image frames. In this study, jitter will be measured in milliseconds (ms), which is how long the data is randomly corrupted. Packet loss occurs when one or more packets of data travelling across a computer network fail to reach their destination. It is typically caused by network congestion but can also have other causes. Packet loss is often measured as a percentage of packets lost with respect to packets sent.

A 10-second section of each of the five reference videos (A, F, G, H, and I as shown in Table B–2) was created in MP4 format as a baseline clip without any induced distortions. This reference clip was then subjected to various levels of distortion resulting in a series of test clips consisting of numerous versions of the same clip, each with a different amount of distortion. These video clips were used for both objective and subjective video quality evaluation.

[10] SysTutorials, “tc-netem (8) – Linux Man Pages.”

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

B.3.3 Video Impairment Process

The video source was streamed using the VLC media player (VLC) on a Dell Latitude E6540 laptop. VLC is a popular open-source, cross-platform, multimedia player compatible with many platforms. The high-definition multimedia interface (HDMI) output of the laptop was connected to the HDMI input of the C60 codec. An active VTC session was established to permit the reference video to be transmitted from the origination endpoint to the terminating endpoint. The two endpoints were connected through an emulated network in order to provide the necessary network impairments used to create the impaired video series. The resulting video received at the terminating endpoint that went through the network was recorded by the Blackmagic HyperDeck 2 recorder. The HDMI output of the terminating codec was connected to the input of the Blackmagic HyperDeck 2 recorder to permit the recording. Figure B–3 represents the capture and recording process through the test bed.

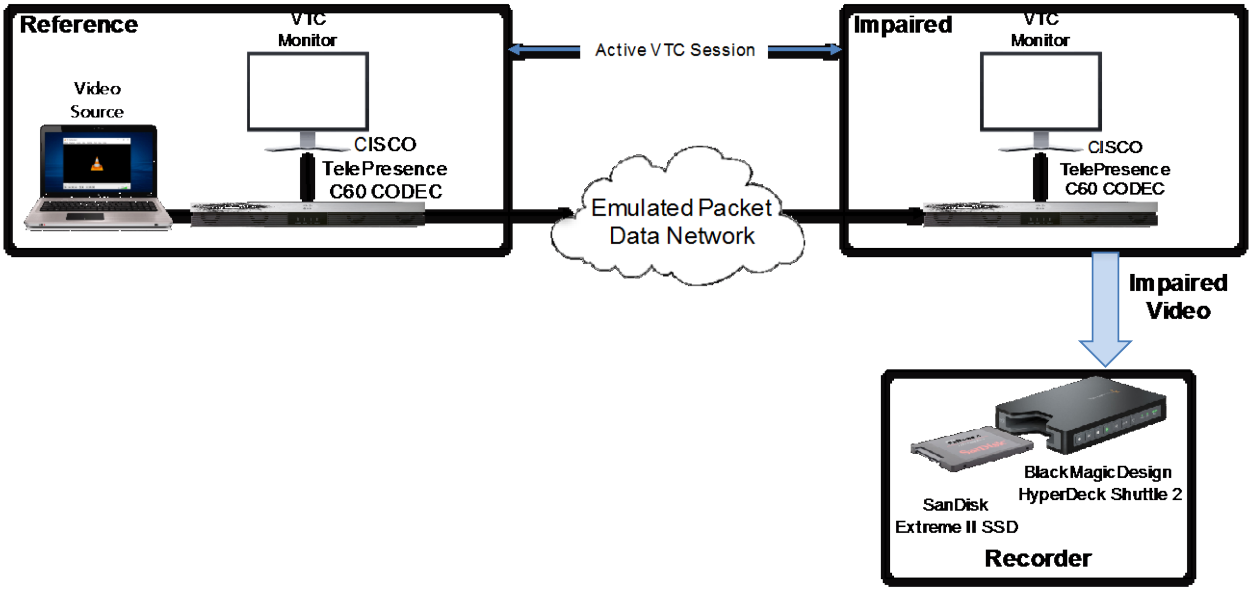


Figure B–3: Recording Impaired Video

The HyperDeck Shuttle is capable of capturing uncompressed 10-bit HD as QuickTime from a Serial Digital Interface (SDI) or HDMI input. However, a direct HDMI to HDMI connection could not be made between the output of the codec and the input of the recorder due to format compatibility issues. As a result, an HDMI to SDI converter was used to connect output of the terminating codec to the input of the HyperDeck recorder to maintain compatibility.

B.3.3.1 Creating Reference Clips from Reference Video

The Opticom tool selected for manipulation of video files can perform PEVQ using video clips ranging in duration from 6 to 20 seconds. However, to comply with subjective video quality assessment methods outlined in the International Telecommunications Union (ITU) Telecommunication Standardization Sector’s (ITU-T) Recommendation P.910,[11] MOSs are

[11] ITU-T Rec. P.910.

limited to a 10-second segment length.[12] Ten-second sections of each reference file were selected to include key scenes of subject interaction and background. Each reference file and corresponding impaired files were trimmed down to 10-second clips and then overlaid with frame numbers. Embedding the frame number in the clip was designed to help synchronize the reference file with the impaired file during the video quality assessment process incorporating the object quality tools (see Appendix A – Measuring Audio and Video Quality).

There are several tools that can be used to trim and overlap frame number on video files. For this study, FFmpeg was selected to process video files and convert between different formats. FFmpeg is a validated Open Source, cross-platform solution for recording, converting, and streaming audio and video.[13] The FFmpeg website[14] contains documentation, descriptions and examples of the Command Line Interface (CLI) language used to trim and stamp the reference video segment.

B.3.3.2 File Standardization

The tools and equipment used to perform full reference objective video assessment need standardized video segments processed with specific characteristics. PEVQ requires specific formats for its input files; as a result, uncompressed Audio Video Interleave (AVI) was selected as the standard for all the video sequences and for all the tools. This format is compatible with the OpenCV tool that was used for SSIM and PSNR video quality assessment.

The recording captured in uncompressed QuickTime format from the recorder was converted to uncompressed AVI using FFmpeg. The FFmpeg website[15] contains documentation, descriptions, and examples of the CLI language used to convert the recorder output into uncompressed AVI.

After the uncompressed AVI was created using FFmpeg from the QuickTime format, VirtualDub was used to trim the video segment as needed. The VirtualDub graphical user interface permits visual feedback allowing precise trimming of the recorded video segment. It also allows the trimmed video to be saved in AVI in its uncompressed state. Software and description were obtained from http://www.virtualdub.org.

Another important aspect of the standardization process is to ensure the reference video file resolution and frame rate matched that of the impaired video file. While it is possible to use FFmpeg to convert resolution and frame rates during the processing stage, it is not recommended due to the possibility of quality loss during the process. Instead, the HDMI output settings of the playback device or camera, and the output setting of the codec were manually configured to ensure that the resolution and frame rate matched at the output of the devices. This will ensure that the reference video and impaired video will have the same resolution and frame rate natively during the recording process. For detailed discussion about correcting frame rate, see Appendix A – Measuring Audio and Video Quality.

[12] ITU-T Rec. P.910.
[13] FFmpeg.
[14] FFmpeg.
[15] FFmpeg.

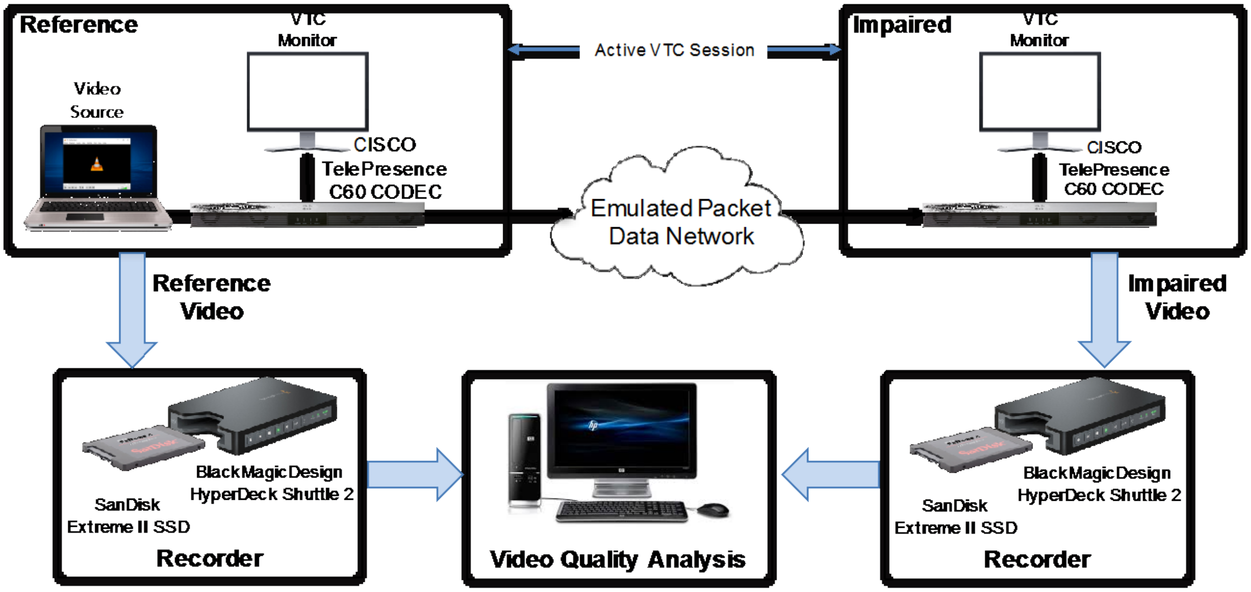


Figure B–4: VTC Testing Environment

B.4 RESULTS

Because the subjective tests would require human participants to evaluate the test videos, effort was made to reduce exposure to videos with minimal perceptible difference. The study team observed the distortion process in real-time in order to ensure that the selected levels encompassed a broad range, from imperceptible levels to high levels of impairment resulting in the distortion values shown in Table B–3. Ten different levels of jitter and ten different levels of packet loss were chosen resulting in the following ranges:

- Jitter: ranged from 0 – 200 msec, at 20 msec intervals
- Packet loss: ranged from 0 – 50%, at 5% intervals

Also included in each series of video clips was a control video clip with no applied jitter or packet loss, a clip with jitter rate of 0, and a clip with a packet loss of 0.

Table B–3: Levels of Distortion for Video Clips

Distortion Levels	
Jitter Rate (ms)	Packet Loss (%)
0	0
20	5
40	10
60	15
80	20
100	25
120	30

Distortion Levels	
Jitter Rate (ms)	**Packet Loss (%)**
140	35
160	40
180	45
200	50

There were five reference video clips with 22 levels of distortion each (11 jitter and 11 packet loss). Additionally, because the test bed applies distortion randomly, the distortion process was performed twice on reference Clip A to allow comparison of two runs of the distortion process (series A and series Z) against a controlled starting video (reference Clip A). This process resulted in six reference clips, each with 22 distinct versions, for a total of 138 total clips.

The reference video clip and the resulting impaired video clips, recorded and standardized into uncompressed formats using the test bed illustrated in Figure B–4, were used as inputs to the objective video quality assessment tools PEVQ, PSNR, and SSIM, as well as subjective testing by human participants.

APPENDIX C. VIDEO QUALITY – OBJECTIVE ASSESSMENT

C.1 OBJECTIVES

One of the main goals of this study was to determine whether any objective measures existed that could be used in an automated fashion to measure quality of VTC video that is collected in courtroom settings. This section outlines the study team's approach for finding and using existing algorithms with this goal in mind.

To accomplish this study objective, the team sought to answer three main questions:

1. Is there a difference between the jitter and packet loss scores?
2. Is there a difference in scores across the five different video Series?
3. Is there a difference between the three Objective Measures (PEVQ™, PSNR, and SSIM)?

C.2 METHODOLOGY

C.2.1 Limitations

The study team chose three video processing algorithms. The team does not claim to have performed an in-depth study of the process, although the algorithms do have a lot of support in the video processing community at large.

C.2.2 Selection of Tools

Based upon the Literature Review and prior experience at JHU/APL, the study team selected three objective tools: SSIM, PSNR, and PEVQ.

PSNR is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed in terms of the logarithmic decibel (dB) scale. PSNR is most commonly used to measure the quality of reconstruction of lossy compression codecs (e.g., for image compression). The signal in this case is the original data, and the noise is the error introduced by compression. When comparing compression codecs, PSNR is an approximation to human perception of reconstruction quality. Although a higher PSNR generally indicates the reconstruction is of higher quality, in some cases it may not. Users have to be extremely careful with the range of validity of this metric; it is only conclusively valid when it is used to compare results from the same codec (or codec type) and same content.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

SSIM is a perceptual metric that quantifies image quality degradation caused by processing, such as data compression or losses in data transmission. It is a full reference metric that requires two images from the same image capture—a reference image and a processed image. SSIM is well known in the video industry, but also has strong applications for still photography. Any image may be used, including those of Imatest LLC test patterns such as Spilled Coins or Log F-Contrast (http://www.imatest.com/).

PEVQ is an end-to-end measurement algorithm to score the picture quality of a video presentation by means of a five-point MOS. It is therefore a video quality model. An independent third party assessed PEVQ's performance during standardization benchmarks by the Video Quality Experts Group (VQEG). Based on the performance results, in which the accuracy of PEVQ was tested against ratings obtained by human viewers, PEVQ became part of the International Telecommunication Union (ITU) Telecommunication Standardization Sector (ITU-T) Recommendation J.247 in 2008.[1]

PEVQ is a full-reference perceptual measurement algorithm that performs pixel analysis of corresponding frames within two videos to generate an assessment of the perceptual quality of the output video. Degradations and artifacts resulting from coding of the video for network transmission are assessed using models of human visual perception. Results of these analyses are converted into a MOS, which have been benchmarked against subjective assessments from human subject testing. In addition, PEVQ records other indicators including PSNR and lip-sync delays.

See Appendix A – Measuring Audio and Video Quality for more details on the application of these tools.

C.2.3 Materials and Methods

Testing was performed at JHU/APL. Objective testing leveraged the test configuration represented in Figure C–1. Equipment was previously acquired for another project and expanded to meet the needs of this project.

[1] ITU-T Rec. J.247.

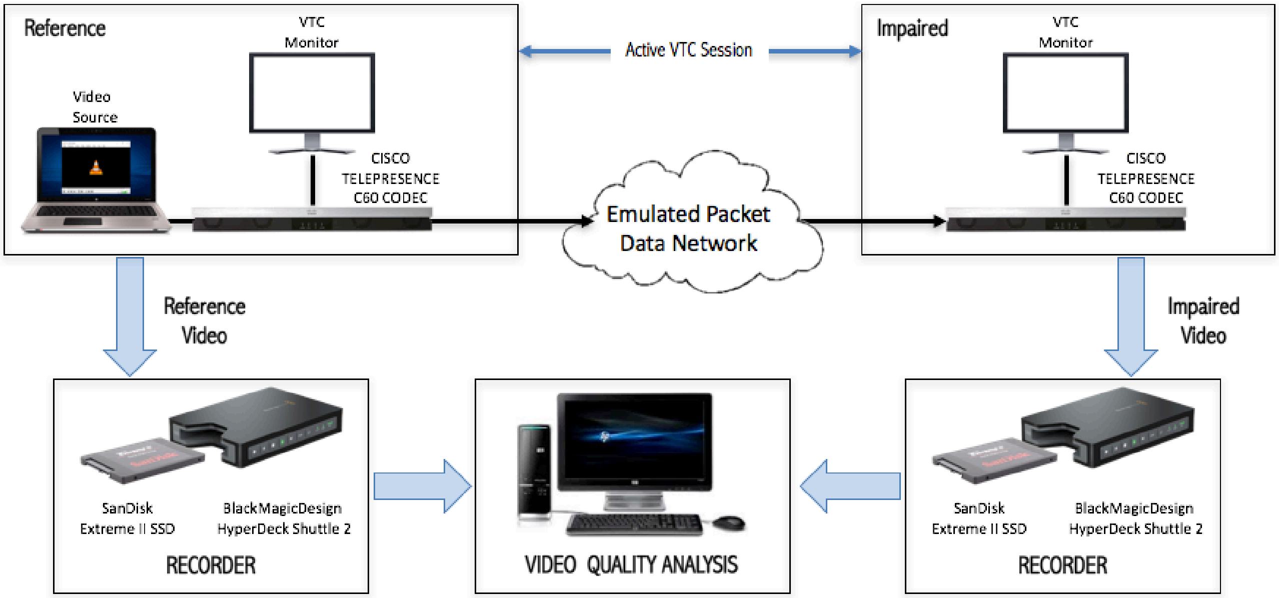


Figure C–1: VTC Testing Environment

Video quality measurement that compares distorted video to the original (reference) video is called full reference video quality assessment. The tools and equipment used to perform full reference objective video assessment require standardized video segments processed with specific characteristics.

For this study, the Opticom PEVQ tool was used to obtain PEVQ mean opinion score (MOS) measurements and the RT&E Center implemented a version of SSIM utilizing the Open Source Computer Vision Library (OpenCV). The VTC testing environment captured the distortions in both the codec and the network. While the testbed inherently captured artifacts from the coding/decoding engine, it was not designed to specifically gauge the performance of the codec. Moreover, the CISCO C60 VTC system used in the test environment was preconfigured with parametric values optimized by the original equipment manufacturer (OEM) to work under various conditions rather than offering multiple options to control the codec engine fine detail. For example, the resolution and frame rate can be adjusted to have video with high resolution over smoothness or smoothness over high resolution. This type of option is designed to mitigate the effects of periodic network degradation, instead of fine-tuning codec engine performance.

C.2.4 Data Collection

Software used to compute PSNR and SSIM during the execution of this study were downloaded from OpenCV[2]. This open source computer vision and machine learning software library includes a comprehensive set of over 2,500 optimized computer vision and machine learning algorithms. The software is freely available under the terms and conditions of a Modified Berkley.

[2] For more information about OpenCV, visit http://opencv.org/about.html.

OpenCV is designed to run on a variety of platforms and operating systems.[3] The VTC project selected two unique platforms (see Table C–1) for purposes of comparison and validation, each sufficient to perform the calculations outlined in the following sections (Table C–1).

Table C–1: Specifications for Two Different Platforms for Computing Objective Video Quality Measurements

Platform	MacBook Pro	Dell Latitude E6530
OS	OS X Version 10.11.6	Windows 7 Enterprise
System Type	64-bit OS	64-bit OS
Processor	Intel Core i7 @ 2.8 GHz	Intel Core i7-3740QM @ 2.7 GHz
Memory	16 GB 1600 MHz DDR3	8 GB
Display Adapter	AMD Radeon R9 M370X 2048 MB	NVIDIA NVS 5200M 1024 MB
Compiler	Apple LLVM 9.0.0	LLVM 5.0.0

C.3 RESULTS – DATA

The data collected during Objective testing with PEVQ, PSNR, and SSIM are included below.

[3] Please refer to the OpenCV Wiki for detailed information and instructions on the proper installation and configuration of the software library, https://github.com/opencv/opencv/wiki.

Table C–2: Objective Metric Results for Sequence A

Clip ID	Jitter (ms)	PEVQ MOS	PSNR Avg (dB)	SSIM Avg	Clip ID	Packet Loss (%)	PEVQ MOS	PSNR Avg (dB)	SSIM Avg
A37	None	5.00	31.45	0.92					
A51	0	5.00	31.47	0.92	A22	0	5.00	31.46	0.92
A62	20	4.86	31.35	0.92	A63	5	4.66	31.20	0.92
A50	40	4.86	31.36	0.92	A58	10	4.67	31.34	0.92
A41	60	4.85	31.30	0.92	A38	15	4.58	31.21	0.92
A72	80	4.79	31.27	0.92	A46	20	4.56	31.32	0.92
A10	100	4.29	31.20	0.92	A89	25	4.55	31.29	0.92
A80	120	2.65	29.06	0.88	A18	30	4.44	31.13	0.92
A45	140	2.02	26.79	0.87	A34	35	4.36	31.30	0.92
A13	160	2.19	28.01	0.86	A69	40	4.26	31.26	0.92
A53	180	2.27	28.51	0.87	A42	45	4.27	31.04	0.92
A28	200	2.40	28.15	0.87	A30	50	3.87	31.14	0.92

Table C–3: Objective Metric Results for Sequence F

Clip ID	Jitter (ms)	PEVQ MOS	PSNR Avg (dB)	SSIM Avg	Clip ID	Packet Loss (%)	PEVQ MOS	PSNR Avg (dB)	SSIM Avg
F88	None	5.00	31.42	0.92					
F84	0	5.00	31.41	0.92	F16	0	4.98	31.28	0.92
F29	20	4.81	31.24	0.92	F38	5	4.68	31.24	0.92
F69	40	4.84	31.29	0.92	F57	10	4.71	31.25	0.92
F41	60	4.84	31.24	0.92	F65	15	4.63	31.24	0.92
F32	80	4.82	31.26	0.92	F12	20	4.55	31.12	0.92
F89	100	4.23	30.60	0.91	F10	25	4.50	31.15	0.92
F98	120	2.34	27.47	0.87	F53	30	4.43	31.15	0.92
F20	140	1	22.64	0.7811	F60	35	4.2	29.52	0.91849
F93	160	1.85	25.07	0.84744	F76	40	3.81	29.06	0.91112
F68	180	1.53	23.44	0.81113	F39	45	3.97	29.29	0.91614
F54	200	1.05	18.84	0.76523	F56	50	3.82	29.16	0.91507

Table C–4: Objective Metric Results for Sequence G

Clip ID	Jitter (ms)	PEVQ MOS	PSNR Avg (dB)	SSIM Avg	Clip ID	Packet Loss (%)	PEVQ MOS	PSNR Avg (dB)	SSIM Avg
G26	None	4.22	31.75	0.90					
G98	0	4.18	31.70	0.90	G29	0	4.21	31.75	0.90
G32	20	4.13	31.58	0.90	G37	5	3.52	31.09	0.89
G74	40	4.11	31.54	0.90	G54	10	4.04	31.54	0.89
G50	60	4.12	31.54	0.90	G73	15	4.02	31.51	0.89
G41	80	4.09	31.56	0.90	G48	20	3.99	31.49	0.89
G46	100	3.72	31.29	0.89	G33	25	3.94	31.45	0.89
G94	120	2.18	28.82	0.84	G81	30	3.93	31.48	0.89
G24	140	2.21	25.69	0.80	G70	35	3.90	31.41	0.89
G86	160	1.00	25.34	0.79	G69	40	3.85	31.40	0.89
G55	180	1.75	26.14	0.81	G96	45	3.84	31.39	0.89
G75	200	1.08	26.83	0.80	G45	50	3.66	31.21	0.89

Table C–5: Objective Metric Results for Sequence H

Clip ID	Jitter (ms)	PEVQ MOS	PSNR Avg (dB)	SSIM Avg	Clip ID	Packet Loss (%)	PEVQ MOS	PSNR Avg (dB)	SSIM Avg
H40	None	4.97	28.96	0.91					
H77	0	4.91	28.65	0.91	H12	0	4.91	28.73	0.91
H67	20	4.99	28.89	0.91	H54	5	4.81	28.84	0.91
H58	40	4.39	28.74	0.91	H95	10	4.60	28.83	0.91
H52	60	3.70	28.33	0.90	H88	15	4.40	28.69	0.91
H38	80	4.98	28.92	0.91	H80	20	4.32	28.66	0.91
H66	100	3.22	27.64	0.89	H69	25	4.24	28.50	0.91
H98	120	1.05	26.29	0.85	H89	30	4.04	28.63	0.91
H26	140	1.00	21.28	0.83	H71	35	3.91	28.59	0.91
H37	160	1.00	22.10	0.80	H53	40	3.86	28.55	0.91
H97	180	1.00	18.78	0.80	H86	45	3.71	28.51	0.91
H48	200	1.00	17.63	0.77	H49	50	2.95	27.26	0.90

Table C–6: Objective Metric Results for Sequence I

Clip ID	Jitter (ms)	PEVQ MOS	PSNR Avg (dB)	SSIM Avg	Clip ID	Packet Loss (%)	PEVQ MOS	PSNR Avg (dB)	SSIM Avg
I28	None	4.25	28.01	0.90					
I39	0	4.31	27.43	0.89	I80	0	4.15	28.21	0.90
I25	20	4.13	28.20	0.90	I34	5	3.97	28.11	0.90
I38	40	3.84	27.99	0.90	I67	10	3.81	28.09	0.90
I55	60	4.17	28.18	0.90	I37	15	3.70	28.06	0.90
I45	80	4.11	28.12	0.90	I23	20	3.59	27.95	0.90
I89	100	2.55	27.19	0.90	I69	25	3.42	27.86	0.90
I87	120	1.13	23.61	0.80	I78	30	3.19	27.73	0.90
I47	140	1.85	25.64	0.88	I42	35	3.05	27.65	0.90
I72	160	1.00	19.96	0.76	I70	40	3.00	27.50	0.90
I59	180	1.00	21.48	0.82	I68	45	2.61	27.21	0.90
I49	200	2.40	28.15	0.87	A30	50	3.87	31.14	0.92

Table C–7: Objective Metric Results for Sequence Z

Clip ID	Jitter (ms)	PEVQ MOS	PSNR Avg (dB)	SSIM Avg	Clip ID	Packet Loss (%)	PEVQ MOS	PSNR Avg (dB)	SSIM Avg
Z70	None	5.00	31.42	0.92					
Z83	0	5.00	31.41	0.92	Z53	0	4.98	31.28	0.92
Z45	20	4.81	31.24	0.92	Z96	5	4.68	31.24	0.92
Z80	40	4.84	31.29	0.92	Z47	10	4.71	31.25	0.92
Z76	60	4.84	31.24	0.92	Z17	15	4.63	31.24	0.92
Z73	80	4.82	31.26	0.92	Z61	20	4.55	31.12	0.92
Z10	100	4.23	30.60	0.91	Z15	25	4.50	31.15	0.92
Z69	120	2.34	27.47	0.87	Z75	30	4.43	31.15	0.92
Z28	140	3.72	30.72	0.91	Z54	35	4.25	30.86	0.92
Z92	160	3.59	30.59	0.91	Z29	40	4.28	30.97	0.92
Z48	180	1.09	25.84	0.84	Z79	45	4.23	31.06	0.92
Z35	200	1.47	27.01	0.85	Z68	50	3.03	29.62	0.89

C.4 DATA ANALYSIS / DISCUSSION

As shown earlier in Table B–3, there were 11 levels of jitter, 11 levels of packet loss, and one sequence with no jitter or packet loss, resulting in a total of 23 measures for each sequence set (A, F, G, H, I, and Z as shown in Table B–2. For the results of each of the objective metrics over each of the video sequences, please see Appendix A – Measuring Audio and Video Quality.

C.4.1 Study Objective #2

To consider whether the tools exhibited a difference in ability to measure quality changes resulting from jitter compared to packet loss, a Pearson correlation analysis was performed among the 11 levels of jitter as well as the 11 levels of packet loss for each of the Objective Metrics and each of the sequences. Table C–8 and Table C–9 show the correlation coefficients for the jitter and packet loss, respectively.

Jitter scores were consistently more correlated (negatively) with the jitter values, whereas the packet loss scores were less correlated with their respective values. This may be due to the innate self-corrections within the VTC system used.

As can be seen in the jitter correlations (Table C–8), there was a strong negative correlation for all of the data points (|coefficient| >.7), indicating that as the jitter increased, the Objective Measures decreased.

Table C–8: Correlation of Jitter vs. Objective Measure for Each Sequence

Jitter (ms)	PEVQ MOS	PSNR Avg	SSIM Avg
Seq. A	-0.90	-0.84	-0.88
Seq. F	-0.91	-0.86	-0.84
Seq. G	-0.92	-0.87	-0.90
Seq. H	-0.91	-0.91	-0.94
Seq. I	-0.92	-0.87	-0.78
Seq. Z	-0.86	-0.74	-0.76

The packet loss correlations (Table C–9), were also negative, but the values, over all, were lower (some showed no correlation at all). Again, the PEVQ MOS has the strongest values, followed by the PSNR, and finally SSIM. In this case, there were some significant differences for the various sequences; in particular, Sequence G had lower values for PEVQ MOS and PSNR.

Table C–9: Correlation of Packet Loss vs. Objective Measure for Each Sequence

Packet Loss (%)	PEVQ MOS	PSNR Avg	SSIM Avg
Seq. A	-0.94	-0.64	0.00
Seq. F	-0.76	-0.72	-0.57
Seq. G	-0.38	-0.37	-0.50
Seq. H	-0.96	-0.67	-0.50
Seq. I	-0.99	-0.97	0.00
Seq. Z	-0.82	-0.67	-0.50

C.4.2 Study Objective #3

The correlation scores were also examined for indication of a difference in scores across the five different video Series. For the jitter values and respective Objective scores, there was very little difference between the five Series (Table C–8 above). For the packet loss values and respective Objective measures, Sequence F had slightly lower correlations and Sequence G had significantly lower correlations with the Objective scores (Table C–9 above).

When the initial objective video quality scores were determined for the Series A videos, each of the objective tools indicated that video quality was relatively steady for jitter rates of 0 through approximately 80 ms, whereupon video quality scores decreased. However, all tools showed an increase in video quality scores around 160 ms of jitter (Figure C–2). To investigate this unexpected trend, a second series of degraded videos was generated from reference video A, called Series Z. Like Series A, Series Z also exhibited steady scores for several levels of jitter before the scores dropped, followed by a rise in scores indicating improved video quality, then a second drop in quality.

Ultimately, jitter scores for each video series showed this pattern of steady initial quality scores, followed by a drop, then a temporary increase in quality scores. The team concluded that this trend indicates that the VTC system compensates for reduced bandwidth, leading to an apparent increase in video quality scores. Video quality scores for packet loss scores were generally much steadier, making the pattern difficult to distinguish.

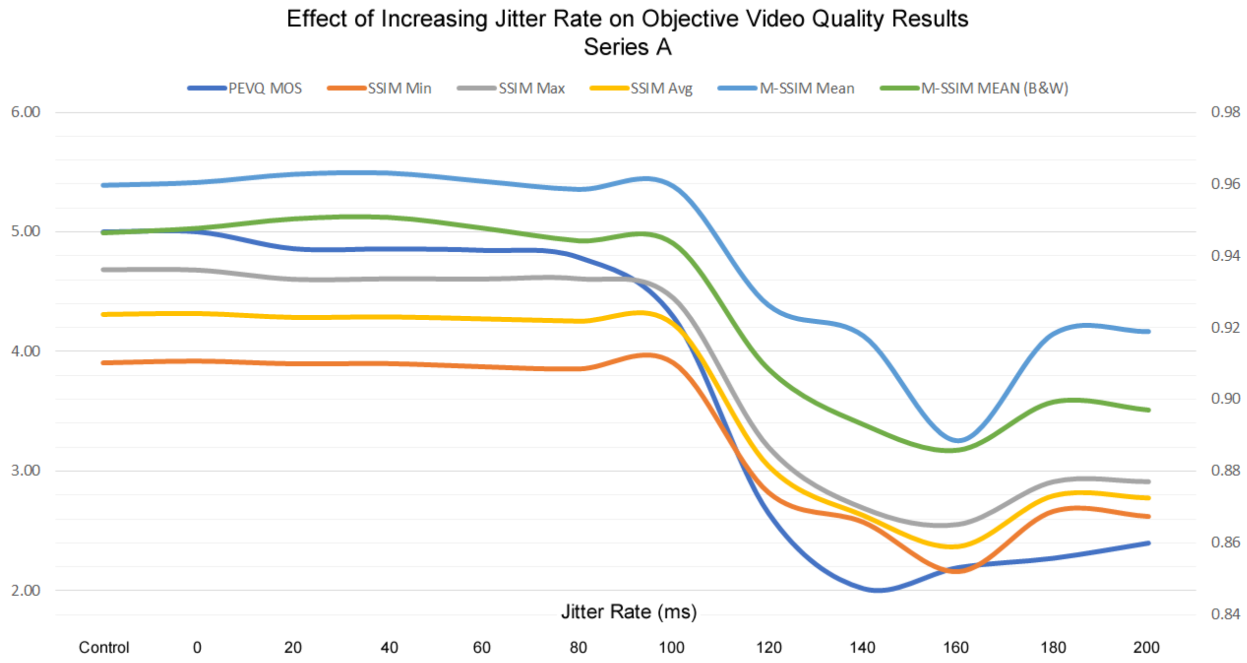


Figure C–2: Unexpected Increase in Video Quality Scores Produced by Objective Tools as Jitter Rate Exceeded 140 ms

C.4.3 Study Objective #3

Following consideration of differences between jitter and packet loss scores, and across the video series, it was possible to consider differences between the three Objective Measures (PEVQ, PSNR, and SSIM).

Of the three Objective Measures, PEVQ had the strongest correlations with jitter and packet loss (with the exception of Sequence G). For jitter, both PSNR and SSIM performed well and it was difficult to rank them against each other, since it was sequence dependent. For packet loss, again PSNR and SSIM performed similarly, however in this case, poorly.

C.5 FINDINGS

Of the three Objective Measures, PEVQ had the strongest correlations with jitter and packet loss (with the exception of Sequence G). For jitter, both PSNR and SSIM performed well and it was difficult to rank them against each other, since it was sequence dependent. For packet loss, again PSNR and SSIM performed similarly, however in this case, poorly.

Jitter scores were consistently more correlated (negatively) with the jitter values, whereas the packet loss scores were less correlated with their respective values. This may be due to the innate self-corrections within the VTC system used.

For the jitter values and respective Objective scores, there was very little difference between the five Series. For the packet loss values and respective Objective measures, Sequence F had slightly lower correlations and Sequence G had significantly lower correlations with the Objective scores.

APPENDIX D. VIDEO QUALITY – SUBJECTIVE ASSESSMENT

In order to compare the performance of the three video quality assessment tools (objective measures), it is important to characterize the same video clips using human observers to assess video quality (subjective measures). The human observer video quality ratings can be used as a benchmark for the tools. This section of the study investigates the human observers' (hereafter referred to as participants) ratings for both overall video quality and a second rating of 'utility' that is operationally defined.

D.1 OBJECTIVES

The study team identified the following study objectives:

1. Is there a difference between mean VQ ratings and Facial Expression Interpretability (FEI) ratings?
2. Is there a difference between mean jitter ratings and packet loss ratings?
3. Is there a difference in mean rating scores across the five different video series?
4. How do the mean ratings (VQ, FEI) correlate to the levels of jitter and packet loss?

It should be noted that these questions are investigational in nature and as such, are not intended to provide an exhaustive analysis of the subjective data. The main interest in collecting the subjective data is to use it as a benchmark for the objective data. The comparison of the objective and subjective data is found in the section on Objective and Subjective Measures – Correlation.

The following sections of the subjective study describe the details of the experiment (to include the video database, test methodology, and procedure) as well as data analysis / results, conclusions, and recommendations.

D.2 METHODOLOGY

The video clips used during the subjective study were the same video clips used for the objective study. The video series described in Table B–2 as well as six distorted clips derived from reference video Series Z were evaluated during the training session of the study.

A single stimulus methodology was used such that each video (including the reference video) was presented for a fixed duration (10 seconds) and then the subject was asked to give two ratings. The first rating was for overall video quality and the second rating was for the ability to interpret facial expression (a proxy for utility). The experiment utilized a within-subjects design; that is, all participants were exposed to all conditions (5 video Series with 11 levels of jitter distortion as well as 11 levels of packet loss distortion) plus the reference video clips. Each individual participant completed the ratings in a single session. To control for practice and fatigue effects, the video clips were randomized using a random number generator for lists found at the random.org website and organized into 10 blocks of 11 clips each. To control for order effects, participants alternated block order such that even-numbered participants completed blocks 1–10 in order while odd-numbered participants began with blocks 6–10 and then completed blocks 1–5. Participants typically completed the entire session in 40 minutes; no more than 30 minutes were spent viewing the video clips in the session.

The overall subjective Video Quality (VQ) rating was aligned to ITU-T MOS definitions. MOS is a widely accepted standard for the subjective assessment of video. It utilizes a continuous five-point Likert-type scale, corresponding to the following adjectives: excellent (5), good (4), fair (4), poor (2), and bad (1).

The second rating of utility allows the participant to assess their ability to identify what is happening, assess emotions and motivations of the speaker, and derive any additional information that would enable them to make judgments regarding the trustworthiness of the speaker and the veracity of his or her testimony. For this study, the team narrowed the scope of the utility rating to the participant's ability to interpret facial expressions, or FEI. It utilizes a continuous five-point Likert-type scale, corresponding to the range of 'cannot at all interpret' (1) through 'can very easily interpret' (5) the facial expression.

For all analyses, Microsoft Excel (version 14.0.6129.5000) with Analysis ToolPak add-in program was used. The Analysis ToolPak is an Excel add-in program that provides data analysis tools for financial, statistical, and engineering data analysis.

D.2.1 Participants

All participants were recruited from the employee distribution list of the Asymmetric Operations Sector of JHU/APL, over the course of two weeks. Prior to participant recruitment, the study team obtained approval through the JHU/APL Human Protections Administrator, who determined from the experiment's test plan that the participants were not considered to be Human Subjects under the Common Rule. Participants did not receive compensation but were able to charge their time to the study contract (approximately 1 hour each).

The 16 participants consisted of a mix of 6 male and 10 female employees who reported corrected visual acuity of at least 20/70 in one eye (the minimum for driving a vehicle in the State of Maryland). All participants except one reported inexperience with video quality assessments; one had prior experience working with drone imagery. Please refer to Table D–1 for a breakout of participant gender, age decade, race/ethnicity, use of corrective lenses, etc.

Table D–1: Summary of Participant Demographics

Demographic Data				
Gender		Wears Glasses		
Male	Female	Yes	No	
6	10	10	6	
Ethnicity		Experience w/ Image Quality		
Hispanic / Latino	Not Hispanic / Latino	Yes	No	
0	16	1	15	
Age Decile				
20's	30's	40's	50's	60's
2	7	1	5	1
Race				
Caucasian	African-American	Asian	Native-American	Other
13	3	0	0	0

D.2.2 Test Session Procedure

The experiment was conducted using a DELL personal computer (PC) laptop in a docking station with a separate 24-inch liquid-crystal display (LCD) monitor and keyboard. The display had a screen resolution of 1920 x 1080 p at 60 Hz. (Dell E2414H flat panel LCD display, 24-inch active area – measured diagonally). The workstation was placed in an office environment with normal indoor illumination levels and an ambient temperature measured at 71 °F. The blinds were drawn to prevent glare on the screen. Subjects viewed the video clips from an approximate distance of 20 inches.

To view the video clips, VLC Media Player (version 2.2.6 Umbrella) was utilized. The video image measured approximately 14 inches wide by 8 inches high (1280 X 720 p) on the screen.

The experiment was conducted in a single session for each participant. Each session included the full set of five video Series with all levels of induced distortion (jitter and packet loss).

After giving informed consent and completing a short demographic questionnaire, each participant was individually briefed about the goal of the experiment and given a demonstration of the experimental procedure. The investigator read from a script to ensure that each participant received the same information.

Participants were presented with a series of short video clips (10 seconds each). After each clip was presented, participants gave two ratings, one for overall video quality and one for the ability to interpret facial expressions. Participants saw repeat videos during the test session. In cases where there were two people shown in the video, participants were instructed to rate the FEI of the person on the right.

The ratings were given on a continuous Likert-type scale of 1 to 5. Ratings could be given up to one decimal place, such as 3.6. The investigator captured the ratings as the participant said them aloud.

For the first rating, overall video quality (based on ITU-T MOS), the numbers corresponded to:

1 = bad
2 = poor
3 = fair
4 = good
5 = excellent

Participants observed the overall video quality scale to use as an aid to rating.

For the second rating, ability to interpret facial expressions, the numbers corresponded to:

1 = cannot at all interpret
2 = cannot interpret
3 = neutral
4 = can interpret
5 = can very easily interpret facial expression

To be clear, the study team did not capture what the facial expression was (such as happiness or disgust), just the ability to determine the facial expression. Participants observed the FEI scale to use as an aid to rating.

Participants gave the overall rating first, followed by the facial expression rating. They were encouraged to take as long as necessary to give the ratings, but to go with their gut instinct. They could take a break at any time but had opportunities to break between each block of 11 video clips. To view the next video clip, participants pressed the ‘n’ on their keyboard for Next. There were 10 blocks of 11 video clips. In addition, each block began and ended with a 2-second blank (black) video clip.

Prior to the test blocks, a training block showing the range of video quality for a training video series was presented to each subject. After each 10-second video clip was shown, the participant gave the two ratings (overall VQ and FEI). After an opportunity to ask questions, the participant was free to begin the experimental session.

After all 10 testing blocks were completed, the participants were thanked and debriefed on the study. They were able to ask any additional questions.

D.3 RESULTS / DATA

The data tables showing the raw data for VQ and FEI ratings by trial, by participant, and descriptive statistics collected during Subjective testing are included in Appendix E – Subjective Assessment Raw Data – VQ and FEI Ratings by Trial, by Participants, and Descriptive Statistics.

D.4 DATA ANALYSIS / DISCUSSION

The raw data for the VQ and FEI ratings by trial by participant are given in Appendix E – Subjective Assessment Raw Data – VQ and FEI Ratings by Trial, by Participants, and Descriptive Statistics, as are the descriptive statistics (mean, standard deviation, min and max) for each trial. There was no missing data.

With regard to the study objectives, the following analyses were conducted and results obtained.

D.4.1 Study Objective #1

To examine evidence of a statistically significant difference between mean VQ ratings and FEI ratings, an F-Test was conducted to test for differences between mean variances, see Table D–2.

Table D–2: F-Test: Two-Sample for Variances, VQ and FEI

	VQ	FEI
Mean	3.284076	3.709748
Variance	1.708419	1.218
Observations	115	115
df	114	114
F	1.402642	
P(F<=f) one-tail	0.036079	
F Critical one-tail	1.362605	

Since F (1.402642) is greater than F Critical one-tail (1.362605, p=0.05), the two variances are unequal. A t-Test to determine if the means are different was conducted, see Table D–3.

Table D–3: t-Test: Two-Sample Assuming Unequal Variances, VQ and FEI

	VQ	FEI
Mean	3.284076	3.709748
Variance	1.708419	1.218
Observations	115	115
Hypothesized Mean Difference	0	
df	222	
t Stat	-2.66843	
P(T<=t) one-tail	0.004091	
t Critical one-tail	1.651746	
P(T<=t) two-tail	0.008183	
t Critical two-tail	1.970707	

The FEI average rating (3.71) is statistically significantly greater than the VQ average rating (3.28).

D.4.2 Study Objective #2

To determine if there was a statistically significant difference between mean jitter ratings and packet loss ratings, an F-Test was conducted to test for differences between mean variances, see Table D–4.

Table D–4: F-Test: Two-Sample for Variances, Jitter and Packet Loss

	Jitter	***Packet Loss***
Mean	3.253646	3.762294
Variance	2.234803	0.58019
Observations	120	110
df	119	109
F	3.851847	
P(F<=f) one-tail	3.3E-12	
F Critical one-tail	1.364678	

Since F (3.851847) is greater than F Critical one-tail (1.364678, p = 0.05), the two variances are unequal. A t-Test to determine if the means are different was conducted, see Table D–5.

Table D–5: t-Test: Two-Sample Assuming Unequal Variances, Jitter and Packet Loss

	Jitter	***Packet Loss***
Mean	3.253646	3.762294
Variance	2.234803	0.58019
Observations	120	110
Hypothesized Mean Difference	0	
df	180	
t Stat	-3.29032	
P(T<=t) one-tail	0.000602	
t Critical one-tail	1.653363	
P(T<=t) two-tail	0.001204	
t Critical two-tail	1.973231	

The packet loss average rating (3.76) is statistically significantly greater than the jitter average rating (3.25).

D.4.3 Study Objective #3:

To look for indication of a statistically significant difference in mean rating scores across the five different video Series, an analysis of variance (ANOVA) was run to test for differences among variances of the five video Series (A, F, G, H, I), see Table D–6 and Table D–7.

Table D–6: Summary of Means and Variances for Video Series (A, F, G, H, I)

Groups	Count	Sum	Average	Variance
a	46	174.2563	3.788179	1.334183
f	46	163.825	3.561413	1.43865
g	46	168.6813	3.666984	1.306426
h	46	145.7454	3.168379	1.622911
i	46	151.7819	3.299606	1.672286

Table D–7: Single Factor ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	12.18012	4	3.045031	2.06458	0.086366	2.411768
Within Groups	331.8505	225	1.474891			
Total	344.0306	229				

Since F (2.06458) is not greater than F Critical (2.411768, p=0.05), the null hypothesis that the mean variances of all five video Series are equal cannot be rejected.

D.4.4 Study Objective #4:

To identify any statistically significant differences related to how the mean ratings (VQ, FEI) correlate to the levels of jitter and packet loss, correlation coefficients were run on the mean ratings (VQ, FEI) and levels of jitter and packet loss, see Table D–8.

As expected, there are strong negative correlations between levels of jitter and packet loss and the corresponding VQ and FEI ratings. The negative correlation simply means that as the level of video distortion goes up, the VQ and FEI ratings go down. However, for Series G, the correlation between packet loss level and VQ and FEI is not as strong. There are strong positive correlations between VQ and FEI ratings.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Table D–8: Correlation Coefficients on the Mean Ratings (VQ, FEI) and Levels of Jitter and Packet Loss (Series A, F, G, H, I)

Series A: Jitter					**Series A: Packet Loss**			
	Level	VQ	FEI			Level	VQ	FEI
Level	1.00				Level	1.00		
VQ	-0.89	1.00			VQ	-0.94	1.00	
FEI	-0.90	0.99	1.00		FEI	-0.93	0.97	1.00
Series F: Jitter					**Series F: Packet Loss**			
	Level	VQ	FEI			Level	VQ	FEI
Level	1.00				Level	1.00		
VQ	-0.92	1.00			VQ	-0.78	1.00	
FEI	-0.92	0.97	1.00		FEI	-0.82	0.94	1.00
Series G: Jitter					**Series G: Packet Loss**			
	Level	VQ	FEI			Level	VQ	FEI
Level	1.00				Level	1.00		
VQ	-0.92	1.00			VQ	-0.65	1.00	
FEI	-0.91	0.97	1.00		FEI	-0.51	0.96	1.00
Series H: Jitter					**Series H: Packet Loss**			
	Level	VQ	FEI			Level	VQ	FEI
Level	1.00				Level	1.00		
VQ	-0.89	1.00			VQ	-0.91	1.00	
FEI	-0.93	0.98	1.00		FEI	-0.83	0.91	1.00
Series I: Jitter					**Series I: Packet Loss**			
	Level	VQ	FEI			Level	VQ	FEI
Level	1.00				Level	1.00		
VQ	-0.88	1.00			VQ	-0.91	1.00	
FEI	-0.91	0.98	1.00		FEI	-0.96	0.95	1.00

D.5 FINDINGS AND RECOMMENDATIONS

FEI ratings were consistently higher than VQ ratings (on average they were half a point higher on the Likert scale.) This suggests that even though participants noticed the jitter and packet loss distortions of the video, they were still able to adequately discern facial expressions of the subjects in the video. There was still a level of utility to the video clip, even though the video was noisy, at least up to a point. This might be a consideration when determining an acceptability threshold for automated video quality assessment tools. The FEI rating as a measure of utility seems to have merit and should be explored further.

Video clips that had packet loss distortions were rated consistently higher in terms of VQ and FEI than clips with jitter distortions (on average they were half a point higher on the Likert scale). This suggests that there is something more objectionable about jitter distortion on the subjective experience of video quality and the ability to interpret facial expressions. Since the experience of jitter affects the horizontal line displacement on the video, it makes sense that this would affect the ability to interpret facial expressions more than packet loss distortion due to network congestion. The source of the noise (jitter or packet loss) might be a consideration when determining an acceptability threshold for automated video quality assessment tools.

The study team found that there was no significant difference for the VQ or FEI mean ratings across the five different video Series. This suggests that the participants were not appreciably affected by background contrast, subject skin color, whether the subject wore glasses (potentially obscuring part of the face), or facial orientation when they gave their VQ and FEI ratings. Participants were able to rate VQ and FEI regardless of the content variations of the video that were presented. This human ability to rate video quality regardless of video content may not be found to the same degree in an automated tool set.

It was not surprising to observe a strong negative correlation between both VQ and FEI ratings and the levels of jitter and packet loss. However, the lack of strong correlation for Series G for both VQ and FEI was somewhat surprising. An explanation could be that for Series G, the subject focus was set at a greater standoff distance from the video camera than the other Series. It was also the only Series that showed the focus subject in a partial-face orientation. As a result, the size of the subject was smaller in terms of perceived visual angle and the actual video display resolution, and less of the face was visible to interpret facial expression. The smaller face area would cause distortion levels to have a greater impact on VQ and FEI ratings. This suggests that a court VTC configuration that shows a large standoff distance or the focus subject in profile would be less tolerant of network-induced distortion.

Since the focus of this subjective study was to generate benchmark ratings for the objective study, there was no planned in-depth analysis of the participant data. It should be considered investigational in nature as no results are definitive with the limited participant sample size of 16 and the limited pool of videos. In future work, it would be interesting to explore if there are differences in video quality ratings based on gender of the participant or the apparent race-matching of the participant to the video subject. Additionally, a study into the effect of experience in video quality assessment should be conducted.

Since the subjective study was conjoined with the objective study, there was a predetermined data-set of video clips (the same ones used in the objective study). While there was diversity of background contrast, some apparent race / ethnicity diversity, and the use of glasses, the videos themselves were not representative of a court VTC setting, see Table B–2. Future investigations should systematically select a diverse range of video clips that are more closely aligned with the court VTC setting.

APPENDIX E. SUBJECTIVE ASSESSMENT RAW DATA – VQ AND FEI RATINGS BY TRIAL, BY PARTICIPANTS, AND DESCRIPTIVE STATISTICS

Video Clip	Rating Type	Distortion Type	Distortion Level	Clip ID	Participant ID															
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a	VQ	j	none	8. A37	5	4.5	5	4	5	5	5	5	5	5	5	5	5	4	4	5
a	VQ	j	0	69. A51	5	4.5	5	4	5	5	5	5	5	5	5	5	5	4.5	5	5
a	VQ	j	20	48. A62	5	4.5	5	3.5	5	4	5	5	5	5	5	5	5	5	5	5
a	VQ	j	40	22. A50	5	4.5	5	3.5	5	4	5	5	5	5	5	5	5	5	5	5
a	VQ	j	60	31. A41	5	4.5	5	4	5	5	5	5	5	5	5	5	5	5	5	5
a	VQ	j	80	38. A72	5	4.5	5	3.5	5	5	5	5	4	5	5	5	5	5	5	5
a	VQ	j	100	2. A10	3	3	2	2	3.9	2	2	3	3	3	1.5	4	3	3	3	2.5
a	VQ	j	120	10. A80	2	2	2	2	2.5	2	2	1.5	3	1	1	2	2	2.5	2	1
a	VQ	j	140	61. A45	1	2.8	2	2.5	2	1	1.5	1	1	1	1	1	2	2	2	1
a	VQ	j	160	84. A13	1	2.5	1	1	2.9	1	1	1	1	1	1	1	1.5	1	2	1
a	VQ	j	180	39. A53	2	1.8	2	2	2.5	2	1.5	2	2	1	1	1	2	2	3	1
a	VQ	j	200	74. A28	2	2	2	2.5	2.7	2	1.5	1.5	2	2	1	1	2	2	2	1
a	VQ	p	0	59. A22	5	4.5	5	4	5	4.5	5	5	5	5	5	5	5	4	5	5
a	VQ	p	5	36. A63	4.5	4.5	5	4	5	5	5	5	5	5	5	5	4.5	5	5	5
a	VQ	p	10	11. A58	4	4.4	4	3.5	4.9	3.5	4.5	4.8	5	4	5	5	4.5	4.5	4	3
a	VQ	p	15	107. A38	4	4.5	4	3	5	5	5	5	5	5	5	5	4	4	5	5
a	VQ	p	20	94. A46	5	4.5	5	4	5	4.5	5	4	5	5	5	5	4	4.5	5	5
a	VQ	p	25	75. A89	4.5	4.4	3.5	3	5	4	4.5	3.5	5	4	3	5	3.5	4	5	2.5
a	VQ	p	30	72. A18	4	4.4	3.5	3.5	4.9	4	3.5	4	5	4	4	4	3.5	4	4	4

Video Clip	Rating Type	Distortion Type	Distortion Level	Clip ID	Participant ID															
					1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***
a	VQ	p	35	54. A34	4	3.5	4	3	4	3	3	3.5	4	5	2	4	4	4	4	3
a	VQ	p	40	51. A69	4	3	4	2.5	4.6	3	3	4	3	4	2	3	3	3	4	2
a	VQ	p	45	105. A42	4	3.5	3	3	4.7	2	3.5	3.8	4	3	3	4	3	3.5	4	3
a	VQ	p	50	100. A30	3	3	3	3	3.7	2.5	2.5	3	3	3	3	3	2.5	3	4	3
a	FE	j	none	8. A37	5	4.5	5	4	5	4	5	5	5	5	3	2	5	5	5	5
a	FE	j	0	69. A51	5	4.5	5	4	5	3	5	5	5	5	3	5	5	5	5	5
a	FE	j	20	48. A62	5	4.5	5	4	5	4	5	5	5	5	3	5	5	4.5	5	5
a	FE	j	40	22. A50	5	4.5	5	4	5	3	5	5	5	5	3	5	5	4.5	5	5
a	FE	j	60	31. A41	5	4.5	5	4	5	4	5	5	5	5	3	5	5	4.8	5	5
a	FE	j	80	38. A72	5	4.5	5	4	5	3	5	5	5	5	3	5	5	4.5	5	5
a	FE	j	100	2. A10	5	3.5	3	4	3.8	3	3	3	4	2	3	4	3.5	3	4	3
a	FE	j	120	10. A80	4.5	2.3	2	3	2	2	3	2.5	3	1	1	3	3	2	3	1
a	FE	j	140	61. A45	3.5	2.3	2	2.5	1	1	2	2	2	1	1	1	2	1.5	2	1
a	FE	j	160	84. A13	2	2.8	1	2.5	2.1	1	1	2	3	2	1	1	2	1.5	2	1
a	FE	j	180	39. A53	5	2.2	2	3	2.3	2	1.5	1.5	3	1	1	1	3	1.8	3	1
a	FE	j	200	74. A28	4	2.3	2	3.5	2.3	2	1.5	2.5	3	2	1	1	2.5	2	2	1
a	FE	p	0	59. A22	5	4.5	5	4	5	4	5	5	5	5	3	5	5	5	5	5
a	FE	p	5	36. A63	5	4.5	5	4	5	4	5	5	5	5	3	5	5	4.5	5	5
a	FE	p	10	11. A58	5	4.5	5	4.5	5	3	5	5	5	4	3	5	5	4.5	5	4
a	FE	p	15	107. A38	5	4.5	4	4	5	4	5	5	5	5	3	5	4	4.5	5	5
a	FE	p	20	94. A46	5	4.5	5	4	5	4	5	4.5	5	5	3	5	4	4.5	5	5
a	FE	p	25	75. A89	4	4.4	4	4	5	4	5	5	5	5	3	5	4	5	5	3.5
a	FE	p	30	72. A18	5	4.4	4	4	5	3	4	5	5	5	3	5	4	4.5	5	4

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Video Clip	Rating Type	Distortion Type	Distortion Level	Clip ID	Participant ID															
					1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***
a	FE	p	35	54. A34	5	3.8	4	4	4.5	4	4	4.2	5	5	3	5	4	4	5	4
a	FE	p	40	51. A69	5	3.5	4	3.5	5	3	4	4	4	5	3	4	3	4	5	3
a	FE	p	45	105. A42	5	3.8	3.5	4	4.7	3	4.5	4	5	4	3	5	3.5	4.5	5	4
a	FE	p	50	100. A30	5	3.2	3.5	4	4.4	3	4	3.5	4	3.5	3	3	3.5	4	5	4
f	VQ	j	none	4. F88	5	4.5	5	4.5	5	4.5	5	5	5	5	5	5	5	5	4	5
f	VQ	j	0	42. F84	5	4.5	5	4	5	5	5	4.8	5	5	5	5	5	5	5	5
f	VQ	j	20	4. F29	4	4.4	5	4	4.9	4	4.5	4.8	5	4	4	5	4	4.5	4	5
f	VQ	j	40	65. F69	5	4.5	5	4.5	5	5	5	5	5	5	5	5	4	5	5	4
f	VQ	j	60	34. F41	5	4.5	5	4	5	5	5	5	5	5	5	5	5	5	5	5
f	VQ	j	80	17. F32	5	4.5	5	4.5	5	5	5	5	5	5	5	5	5	4.8	5	5
f	VQ	j	100	25. F89	3	3	3	2.5	4.5	3	2.5	4	3	3	2	4	3.5	2.5	4	2.5
f	VQ	j	120	95. F98	3	2.8	2	2.5	3	1	2	2.5	2	2	1	2	3	2.5	3	2
f	VQ	j	140	20. F20	1	1.5	1	1	1.3	1	1	1	1	1	1	1	1	1	2	1
f	VQ	j	160	109. F93	1	2.5	2	1.5	2.7	1	1	1.5	1	1	1	1	2	2	2	1
f	VQ	j	180	104. F68	1	1.5	2	2	2	1	1	1	2	1	1	1	1	1.5	2	1
f	VQ	j	200	91. F54	1	1.5	1	1.5	1.5	1	1	1	1	1	1	1	1	1	1	1
f	VQ	p	0	13. F16	4	4.5	5	4	5	5	5	4.8	5	5	5	5	5	4.5	5	5
f	VQ	p	5	55. F38	3	3	3	2.5	2.8	3	2	3	4	3	2	3	3.5	3.5	3	2
f	VQ	p	10	83. F57	3	3.8	3.5	3.5	4.7	2.5	3	4.5	5	3.5	4	4	4	3.5	5	3
f	VQ	p	15	16. F65	4	3	4	3.5	4.7	2	3.5	4	4	3	4	5	3.5	4	5	3
f	VQ	p	20	19. F12	4	3.5	3.5	3	4.7	3	3	4.5	4	4	4	4	4	4	4	3
f	VQ	p	25	63. F10	4	3.5	3	3.5	4.7	2	4	4	3.5	4	4	4	3.5	4	4	2
f	VQ	p	30	86. F53	4	3	2.5	3	4	2	2.5	3.5	4	3	3	4	3.5	3	4	2

Video Clip	Rating Type	Distortion Type	Distortion Level	Clip ID	Participant ID															
					1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
f	VQ	p	35	15. F60	3	2.8	4	2.5	3.8	2	2.5	3.8	3	3	2	4	2.5	3.5	3	2
f	VQ	p	40	45. F76	3	4	3	2.5	3.7	3	3	2.5	2	2	3	3	2.5	2.5	4	2
f	VQ	p	45	93. F39	3.5	3	2.5	2	3.9	2	2.5	3	3	3	2	2	3	2.5	4	2
f	VQ	p	50	108. F56	3	3	2.5	2	3.9	2	2	2	2	2	2	2	3	2.5	3	2.5
f	FE	j	none	4. F88	5	4.5	5	4.5	5	5	5	5	5	5	5	2	5	3	5	5
f	FE	j	0	42. F84	5	4.5	5	4.5	5	4	5	5	5	5	3	5	5	5	5	5
f	FE	j	20	4. F29	5	4.5	5	4.5	4.9	4	4	5	5	5	4	5	5	5	5	5
f	FE	j	40	65. F69	5	4.5	5	4.5	5	4	5	5	5	5	3	5	4	5	5	4
f	FE	j	60	34. F41	5	4.5	5	4.5	5	4	5	5	5	5	3	5	5	5	5	5
f	FE	j	80	17. F32	5	4.5	5	4.5	5	4	5	5	5	5	3	5	5	5	5	5
f	FE	j	100	25. F89	5	3.2	4	4	4.8	4	4	4.5	4	4	3	5	4	4	5	3.5
f	FE	j	120	95. F98	5	3	2.5	3.5	3.8	1	3	3	4	4	2	2	4	3.5	5	3
f	FE	j	140	20. F20	2	2	2	2.5	1.5	1	1	1.5	2	1	1	1	2	1.8	3	1
f	FE	j	160	109. F93	3	2.8	2	2.5	1.4	1	1.5	3.5	3	1	1	1	3	2	2	1
f	FE	j	180	104. F68	3	2	2	3	1.2	1	1.5	1.5	4	1	1	1	1	2	2	1
f	FE	j	200	91. F54	3	1.8	1	2	1	1	1.5	1.5	1	1	1	1	1	1.5	2	1
f	FE	p	0	13. F16	5	4.5	5	4.5	5	4	5	5	5	5	4	5	5	5	5	5
f	FE	p	5	55. F38	5	3.5	4	4	3.3	4	3	4	5	4	3	5	4	4.5	4	3.5
f	FE	p	10	83. F57	5	4.2	4	4.5	5	3	4	5	5	5	3	5	4	5	5	4
f	FE	p	15	16. F65	5	3.5	4	4.5	4.9	3	4.5	5	4	4	3	5	4	5	5	4
f	FE	p	20	19. F12	5	4.2	4	4.5	4.8	4	4.5	5	5	5	3	5	4	4.8	5	4
f	FE	p	25	63. F10	5	3.8	4	4.5	5	4	4.5	5	4	4	3	5	4	5	5	3.5
f	FE	p	30	86. F53	5	3.8	3.5	4	4.5	3	4	4.5	5	4	4	5	4	4.5	5	3.8

Video Clip	Rating Type	Distortion Type	Distortion Level	Clip ID	Participant ID															
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
f	FE	p	35	15. F60	5	3.2	4	4	4.7	2	4	4.5	4	4	2	5	3	4.5	5	3
f	FE	p	40	45. F76	5	4.3	4	3.5	4.3	3	4	2.5	4	2	3	4	3.5	3	5	3
f	FE	p	45	93. F39	5	3.2	3	3.5	4.2	2	3.5	4	4	4	3	3	4	3.5	5	3
f	FE	p	50	108. F56	5	3.8	3.5	3.5	4.1	3	3.5	4	4	4	2	3	4	3	4	3.5
g	VQ	j	none	2. G26	4	4.5	5	4	5	4	5	5	5	5	5	5	5	4	5	5
g	VQ	j	0	67. G98	4.5	4.5	5	4.5	5	5	5	5	5	5	5	5	5	5	5	5
g	VQ	j	20	12. G32	5	4.4	5	3.5	5	5	5	5	5	5	5	5	5	5	5	5
g	VQ	j	40	99. G74	5	4.5	5	4	5	5	5	4.5	5	5	5	5	4	4.5	5	5
g	VQ	j	60	58. G50	5	4.5	5	4.5	5	4	5	5	5	5	5	5	4	5	5	4
g	VQ	j	80	76. G41	5	4.5	5	4.5	5	4	5	5	5	5	5	5	5	5	5	4
g	VQ	j	100	77. G46	3	3	3	2.5	3.9	2	2	3.5	3	3	3	4	3.5	3	4	2.5
g	VQ	j	120	21. G94	1	3	2	2	2.9	2	2	1.5	2	1	1	1	2	2	2	2
g	VQ	j	140	41. G24	1.5	2.8	3	1.5	2.5	2	1.5	2	2	1	1	1	2.5	1.5	1	1
g	VQ	j	160	56. G86	1	2.5	1	2	1.5	1	1	1.5	1	2	1	1	1	1.5	1	1
g	VQ	j	180	103. G55	5	1.5	2	2.5	3	1	1	1	2	1	1	1	1	1.5	2	1
g	VQ	j	200	53. G75	1	1.5	2	1.5	1.5	2	1	1	2	1	1	1	2	1	1	1
g	VQ	p	0	89. G29	5	4.4	5	4	5	5	5	5	5	5	5	5	4	5	5	5
g	VQ	p	5	80. G37	3.5	3.5	4	3	3.5	2	1.5	2	3.5	2	3	3	3	3	3	1
g	VQ	p	10	49. G54	5	4.5	5	3.5	5	4	5	5	5	5	5	5	5	4.5	5	5
g	VQ	p	15	35. G73	4.5	3.8	4	4	5	5	5	5	4	5	4	5	5	3.5	4	4
g	VQ	p	20	57. G48	4	4.4	4	4	5	2	5	4	4	3	4	4	4	4	4	4
g	VQ	p	25	98. G33	4	4.4	3	3.5	4.5	2	4	4.5	4	4	4	4	3.5	3	4	4
g	VQ	p	30	5. G81	3	4	3.5	3	4.9	3	4.5	4.2	5	5	4	4	4	3.5	4	3

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Video Clip	Rating Type	Distortion Type	Distortion Level	Clip ID	Participant ID															
					1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
g	VQ	p	35	88. G70	4	3	3.5	3	4.5	2	3	4	3	4	2	4	3	3	4	3
g	VQ	p	40	26. G69	3	3.5	3	2.5	4	3	4	4	4	3	3	2	3.5	3	5	2.5
g	VQ	p	45	85. G96	3	3	3	3	3.8	2	2	4	3	3	2	3	3	3	4	2
g	VQ	p	50	33. G45	4	3.5	2	2.5	3.2	3	2.5	2.5	3	2.5	2	4	3	2	3	2
g	FE	j	none	2. G26	5	4	5	4.5	4	2.5	5	4	5	5	5	5	3.5	4	5	5
g	FE	j	0	67. G98	5	4.5	5	4	4.3	3	5	4	5	5	4	5	5	4.5	5	5
g	FE	j	20	12. G32	5	4.4	5	4	4	4	5	5	5	5	4	5	4	4.5	5	5
g	FE	j	40	99. G74	5	4.5	5	4	4.3	3	5	5	5	5	4	5	4	4.5	5	5
g	FE	j	60	58. G50	5	4.5	5	4	4.3	3	5	4	5	5	4	5	5	4.5	5	5
g	FE	j	80	76. G41	5	4.5	5	4	4.3	3	5	4	5	5	4	5	5	4.5	5	4.5
g	FE	j	100	77. G46	5	3.8	4	4	3.5	3	3.5	4	5	4	4	5	4	4	5	3.5
g	FE	j	120	21. G94	2	3.2	3	3	3.3	2	2.5	2.3	3	1	2	2	2	1.5	3	3.5
g	FE	j	140	41. G24	3	3.2	2	3	2.3	2	2	2	3	1	1	3	3.5	2	3	2.5
g	FE	j	160	56. G86	1.5	2.8	1	2	1	1	1	1	3	1	1	1	1	1.5	1	3.5
g	FE	j	180	103. G55	3	2.3	3	3	2.3	2	2	2	4	1	1	2	2	2	3	3
g	FE	j	200	53. G75	1	2.1	2	2	1	2	1.5	1.5	4	1	1	1	2	1.2	1	2
g	FE	p	0	89. G29	5	4.4	5	4	4.3	3	5	5	5	5	5	5	5	4.5	5	5
g	FE	p	5	80. G37	5	3.8	4	4	3	3	2.5	3.5	5	4	2	4	3.5	3.5	4	2
g	FE	p	10	49. G54	5	4.5	5	4	4.3	3	5	5	5	5	4	5	5	4.5	5	5
g	FE	p	15	35. G73	5	4.2	4	4	4.3	3	5	5	5	5	4	5	5	4.5	5	4.6
g	FE	p	20	57. G48	5	4.4	4	4	4.3	3	5	4	5	4	4	5	4	4	5	5
g	FE	p	25	98. G33	5	4.4	3.5	4	4.2	3	5	5	5	5	4	5	4	4	5	5
g	FE	p	30	5. G81	5	4.5	3.5	4	4	3	4.5	5	5	5	3	5	4	4	5	4

Video Clip	Rating Type	Distortion Type	Distortion Level	Clip ID	Participant ID															
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
g	FE	p	35	88. G70	5	3.8	4	4	4.3	3	4.5	4	4	4	3	5	4	3.5	5	4
g	FE	p	40	26. G69	5	3.8	4	3.5	4	3	4.5	4.5	5	4	3	4	4	3	5	3.5
g	FE	p	45	85. G96	4	3.8	4	4	4	3	3.5	4.5	5	4	3	4	4	4	5	3.5
g	FE	p	50	33. G45	5	3.8	3	3.5	3.7	3	4	4	4	2.5	4	5	3	3	4	3
h	VQ	j	none	10. H40	5	4.5	5	3.5	5	2.5	5	4	5	2	5	5	5	5	5	5
h	VQ	j	0	9. H77	5	4.5	5	3.5	5	5	5	5	4	5	5	5	5	4.5	1	4
h	VQ	j	20	110. H67	5	4.5	5	4	5	5	5	5	5	5	5	5	4	5	5	5
h	VQ	j	40	90. H58	4	4.3	3.5	3	5	2.5	3.5	4	4	4.5	4	5	4	4	4	4
h	VQ	j	60	1. H52	3	3	3	2.5	3.5	2	4	3	3	3	1	4	3	3	3	2.5
h	VQ	j	80	81. H38	5	4.5	5	4	5	4.5	5	5	5	4	5	5	5	5	5	5
h	VQ	j	100	24. H66	3	2	2	2	2.3	2	2.5	2.5	1	1	1	2	2.5	2	3	1
h	VQ	j	120	23. H98	1	1.5	1	1	1.5	1	1.5	2	1	1	1	1	1.5	2	1	1
h	VQ	j	140	43. H26	1	1.5	1	1.5	2.3	2	1	1	1	1	1	1	1.5	1	1	1
h	VQ	j	160	101. H37	1	1.5	1	1.5	1.5	1	1	1	1	1	1	1	1	1	1	1
h	VQ	j	180	8. H97	1	1.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1
h	VQ	j	200	68. H48	1	1.8	1	1	1	1	1	1	1	1	1	1	1	1	1	1
h	VQ	p	0	106. H12	5	4.5	5	4	5	4	5	5	5	5	5	5	5	5	5	5
h	VQ	p	5	52. H54	4	3.5	4	2.5	4.9	3	3.5	4.2	4	4	4	4	3.5	3.5	4	3
h	VQ	p	10	40. H95	3.5	4	3	3	4.4	3	3.5	3.5	3	4	3	4	3	3.8	4	3
h	VQ	p	15	47. H88	4	3.8	3	2.5	4.5	3	4	3.2	3	3	4	4	3	3	3	2
h	VQ	p	20	71. H80	3	3.5	2.5	2.5	4.3	2	2	3	3	3	2	3	3	3	4	2
h	VQ	p	25	60. H69	3.5	3.2	2.5	2.5	4.3	2	2	3	3	3	3	3	3	3	4	3
h	VQ	p	30	18. H89	3	3.8	3	2.5	4.7	2	2.5	3	3	3	1	4	3	3	3	3

Video Clip	Rating Type	Distortion Type	Distortion Level	Clip ID	Participant ID															
					1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
h	VQ	p	35	50. H71	3	3	2	2	3.7	2	2	2.5	2	2	2	3	3	2.5	4	2
h	VQ	p	40	28. H53	3	3.5	2	2	4	2	2	3	3	2	2	3	3	2.5	3	2
h	VQ	p	45	73. H86	2	3.5	2	2.5	4	2	2	3	2	2	2	3	3	3	4	2
h	VQ	p	50	44. H49	2	2	1	1.5	2.5	2	1.5	1	1	1	1	1	1	2	2	1
h	FE	j	none	10. H40	5	4	5	4	4	3	5	5	5	4	4	3	4	3	5	5
h	FE	j	0	9. H77	5	4.5	5	4.5	5	4	5	5	4	5	4	5	4.5	3.5	5	5
h	FE	j	20	110. H67	5	4.5	5	4	5	4	5	5	5	5	4	5	4	3.5	5	5
h	FE	j	40	90. H58	5	4.4	4	4	5	3	4.5	5	5	5	4	5	5	3.5	5	5
h	FE	j	60	1. H52	5	3.5	4	4	3.9	3	4	4.5	3	2	3	5	4	3.5	4	4
h	FE	j	80	81. H38	5	4.5	5	4	5	3	5	5	5	5	5	5	5	4	5	5
h	FE	j	100	24. H66	4.5	2.5	4	3.5	3.2	2	3	3	2	2	2	3	3	2	4	2.5
h	FE	j	120	23. H98	3	2	2	1	1.7	1	2	2	2	2	1	1	2.5	1.5	2	2.5
h	FE	j	140	43. H26	3	2	1	3	2	2	1	1	2	1	1	1	2	1	2	1
h	FE	j	160	101. H37	2	2.3	1	3.5	1.2	1	1.5	1.5	2	1	1	1	2	1	2	3
h	FE	j	180	8. H97	2	2.3	1	2	1	1	1	1	1	1	1	1	1.5	1.2	1	1
h	FE	j	200	68. H48	2	2	1	2	1	1	1	1	1	1	1	1	1	1.5	1	1
h	FE	p	0	106. H12	5	4.5	5	4	5	4	5	5	5	5	4	5	5	3.5	5	5
h	FE	p	5	52. H54	5	3.8	4	4	5	3	4.5	5	5	5	4	5	4	3.5	5	4
h	FE	p	10	40. H95	5	4.2	4	4	4.4	4	4.5	5	4	5	4	5	4	3	5	4
h	FE	p	15	47. H88	5	4.2	4	4	4.5	4	4.5	5	4	4	4	5	4	3	5	4
h	FE	p	20	71. H80	5	3.8	3.5	4	4.7	2	3	5	4	4	2	4	4	3	5	3.5
h	FE	p	25	60. H69	5	3.8	3.5	3.5	5	4	3.5	4	5	4	3	4	4	3.5	5	4
h	FE	p	30	18. H89	5	3.8	4	3.5	4.7	3	4	4.8	4	4	3	5	4	3	4	4.5

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Video Clip	Rating Type	Distortion Type	Distortion Level	Clip ID	Participant ID															
					1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
h	FE	p	35	50. H71	4.5	3.5	3.5	3.5	4.4	3	3	4.2	4	2.5	2	3	3	2	4	4
h	FE	p	40	28. H53	4.5	4.2	4	3.5	4.7	3	4	4	4	2	2		4	2.5	5	3
h	FE	p	45	73. H86	5	3.8	3	3.5	4.3	4	3	4	3	3.5	2	4	3.5	4	5	3.5
h	FE	p	50	44. H49	4	2.3	2	3	2	2	2.5	2	2	1	1	1	1	1	3	2
i	VQ	j	none	6. I28	4	4.7	5	4.5	5	5	5	5	5	5	5	5	5	5	5	4
i	VQ	j	0	96. I39	5	4.5	5	4.5	5	5	5	4.5	5	5	5	5	4	5	5	4
i	VQ	j	20	66. I25	5	4.5	4.5	4.5	5	5	4.5	5	5	4	5	5	4	5	5	4
i	VQ	j	40	27. I38	4	3.8	3.5	3	4.9	3	4	4	5	4	4	4	4	3.2	5	4
i	VQ	j	60	82. I55	5	4.5	5	4.5	5	4.5	5	5	5	5	5	5	5	5	5	5
i	VQ	j	80	6. I45	5	4.5	5	4	5	5	5	5	5	5	2	5	5	4.5	5	5
i	VQ	j	100	102. I89	2	2.3	2	2	3	1	2	2	2	1	1	1	2	2	3	2
i	VQ	j	120	37. I87	1	1.5	1	1	1.8	1	1	1	2	1	1	1	1	1.5	1	1
i	VQ	j	140	32. I47	2	2	2	2	2.3	2	1	1.5	1	1	1	1	2	2	2	1
i	VQ	j	160	29. I72	1	1.5	1	1	1.3	1	1	1	1	1	1	1	1	1	1	1
i	VQ	j	180	79. I59	1	1.5	1	1.5	1.2	1	1	1	1	1	1	1	1	2	2	1
i	VQ	j	200	62. I49	1	2.2	1	1.5	2	1	1	1	1	1	1	1	1	1.5	1	1
i	VQ	p	0	30. I80	5	4.5	5	4.5	5	5	5	5	5	5	5	5	5	5	5	5
i	VQ	p	5	3. I34	4.5	4.5	4	3.5	5	5	5	4.5	5	4	4	5	4	4	4	3.5
i	VQ	p	10	46. I67	4	3.8	3	3	3.7	3	4	4	4	3	4	4	3	3	4	3
i	VQ	p	15	70. I37	4	3.5	3	3	4.5	2	3	3	3	3	2	4	3	3.5	3	2
i	VQ	p	20	7. I23	3	3.8	3	2.5	3.9	3	2.5	3.8	4	2.5	1	4	3	2.5	3	3
i	VQ	p	25	92. I69	3	2.3	3	2.5	3.9	2	2	3.5	2.5	3	2	3	2	2.5	3	3
i	VQ	p	30	87. I78	3	3	2	2	3	2	2	2.5	2	2	2	3	3	2	3	2.9

Video Clip	**Rating Type**	**Distortion Type**	**Distortion Level**	**Clip ID**	**Participant ID**															
					1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***
i	VQ	p	35	78. I42	3	3	2	2	3.7	2	2	2.5	2	2	2	3	3	3	3	2
i	VQ	p	40	64. I70	3	3.2	2.5	2.5	3.7	2	2.5	2.5	3	2	2	3	3	3	3	2
i	VQ	p	45	97. I68	3	2.8	2	2	3.5	1	2	2	2	2	2	2	2	2	3	2
i	VQ	p	50	14. I62	2	2.8	2	2	2.8	2	2	2.5	2	2	1	2	2	2	2	2
i	FE	j	none	6. I28	5	4.7	5	4.5	5	4	5	5	5	5	5	2	5	5	5	4
i	FE	j	0	96. I39	5	4.5	5	4.5	5	4	5	5	5	5	4	5	4	5	5	5
i	FE	j	20	66. I25	5	4.5	5	4.5	5	4	5	5	5	5	4	5	4	5	5	4
i	FE	j	40	27. I38	5	4.2	4	4	5	3	5	5	5	5	3	5	4	4.5	5	4
i	FE	j	60	82. I55	5	4.5	5	4.5	5	3	5	5	5	5	5	5	5	4.5	5	5
i	FE	j	80	6. I45	5	4.5	5	4.5	5	4	5	5	5	5	3	5	5	4.5	5	5
i	FE	j	100	102. I89	4	2.8	3	3.5	3.2	2	3.5	3.5	4	2.5	1	2	3	3	4	4
i	FE	j	120	37. I87	2	2.5	2	2.5	1.8	1	1.5	1.5	3.5	1	1	1	1	1.5	4	2
i	FE	j	140	32. I47	4	2.3	2	3	3	2	2	2	2	1	1	1	3	3	3	1
i	FE	j	160	29. I72	1	2.5	1	1	1.3	1	1	1	1	1	1	1	1	1.2	1	1
i	FE	j	180	79. I59	3	2	1	2	1	1	1	2	1	1	1	1	2	1.5	3	2
i	FE	j	200	62. I49	3	2.2	1	2	1	1	1	2	1	1	1	1	2	2	1	2
i	FE	p	0	30. I80	5	4.5	5	4.5	5	4	5	5	5	5	4	5	4	5	5	5
i	FE	p	5	3. I34	5	4.5	5	4.5	5	4	5	5	5	5	3	5	4	4.5	5	4
i	FE	p	10	46. I67	5	4.2	4	4	4.3	3	4.5	5	4	4	4	4	4	4	5	3.5
i	FE	p	15	70. I37	5	3.8	4	4	4.8	2	4	5	4	4	3	5	4	4.5	5	3.5
i	FE	p	20	7. I23	5	4.2	3.5	4	4.3	3	3.5	5	4	3.5	2	5	4	4	5	4
i	FE	p	25	92. I69	5	2.8	3.5	4	4.7	3	3.5	4.5	5	4	2	4	3.5	3	5	4
i	FE	p	30	87. I78	5	3.8	3	3.5	4	3	3.5	4	4	4	2	3	4	3.5	5	3.9

Video Clip	Rating Type	Distortion Type	Distortion Level	Clip ID	Participant ID															
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
i	FE	p	35	78. I42	5	3.8	3.5	3.5	4.1	3	3.5	4	4	3.5	2	4	4	3.5	4	3
i	FE	p	40	64. I70	5	3.8	3.5	4	4.5	2	3	4	4	4	3	4	3	4	5	3
i	FE	p	45	97. I68	5	3.2	3	3.5	3.8	1	3.5	3.5	4	2	3	3	3	4	4	3.8
i	FE	p	50	14. I62	5	3.2	2.5	3.5	3.7	2	4	3.5	2	1	1	3	3.3	3	4	3.5
				Block order	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2

APPENDIX F. BIBLIOGRAPHY

Acharaya, Tinku, and Ray K. Ray. Image Processing: Principles and Applications. (Hoboken, NH: John Wiley & Sons, 2005). http://onlinelibrary.wiley.com/book/10.1002/0471745790.

Bellone, E. "Private Attorney-Client Communications and the Effect of Videoconferencing in the Courtroom," *The Journal of International Commercial Law and Technology* 8, no. 1 (2013):24–48.

Center for Legal & Court Technology (CLCT). "Best Practices for Using Video Teleconferencing for Hearings and Related Proceedings," *Report to the Administrative Conference of the United States*, November 6, 2014, 14. 39–40

Center for Legal & Court Technology(CLCT). "Best Practices for Using Video Teleconferencing for Hearings and Related Proceedings," *Draft Report to the Administrative Conference of the United States* (Williamsburg, Virginia: 8 October 2014). https://www.acus.gov/sites/default/files/documents/Draft_Best%2520Practices%2520Video%2520Hearings_10-09-14_1.pdf.

Cisco. "Jitter and Network Delay," *Cisco Community*, https://communities.cisco.com/docs/DOC-47410, 24 January 2014.

Cisco Systems. D15029.01 "Troubleshooting Guide TC 6.0," April 2013, https://www.cisco.com/c/dam/en/us/td/docs/telepresence/endpoint/codec-c-series/tc6/troubleshooting_guide/tc_troubleshooting_guide_tc60.pdf.

Côtė, Nicholas. "Integral and Diagnostic Intrusive Perception of Speech Quality," *T-Labs Series in Telecommunications Services*, 1. doi:10.1007/978-3-642-18463-5.

Davis, Robin, Billie Jo Matelevich-Hoang, Alexandra Barton, Sara DebusSherrill, and Emily Niedzwiecki. "Research on Videoconferencing at Post-Arraignment Release Hearings: Phase I Final Report," Fairfax, VA: ICF International (29 May 2015). https://www.ncjrs.gov/pdffiles1/nij/grants/248902.pdf

Diamond, Shari, Locke Bowman, Manyee Wong, and Matthew Patton. "Efficiency and Cost: The Impact of Videoconferenced Hearings on Bail Decisions," *Journal of Criminal Law and Criminology* 100, no. 3 (Summer 2010):869–902.

FFmpeg (web page). "About FFmpeg," accessed 30 October 2017, http://ffmpeg.org/about.html.

Goodman, Gail S., Ann E. Tobey, Jennifer M. Batterman-Faunce, Holly Orcutt, Sherry Thomas, Cheryl Shapiro, and Toby Sachsenmaier. "Face-to-face Confrontation: Effects of Closed Circuit Technology on Children's Eyewitness Testimony," *Law and Human Behavior* 22, no. 2 (April 1998): 165–203. https://link.springer.com/article/10.1023/A:1025742119977.

Haas, Aaron. “Videoconferencing in Immigration Proceedings,” *University of New Hampshire Law Review* 5 no. 1 (December 2006):57–90. http://scholars.unh.edu/cgi/viewcontent.cgi?article=1084&context=unh_lr.

Huynh-Thu , Quan, Marie-Neige Garcia, Filippo Speranza, Philip Corriveau, and Alexander Raake. “Study of Rating Scales for Subjective Quality Assessment of High-Definition Video,” in *IEEE Transactions on Broadcasting* 57, no. 1 (March 2011):1–14. doi:10.1109/TBC.2010.2086750.

ITU Telecommunication Standardization Sector (ITU-T). Recommendation P.861, “Objective quality measurement of telephone-band (300–3400 Hz) speech codecs,” August 1996, https://www.itu.int/rec/T-REC-P.861-199608-S/en.

ITU Telecommunication Standardization Sector (ITU-T). Recommendation P.862, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” February 2001, https://www.itu.int/rec/T-REC-P.862-200102-I/en.

ITU Telecommunication Standardization Sector (ITU-T). Recommendation P.863, “Perceptual objective listening quality prediction,” March 2018, https://www.itu.int/rec/T-REC-P.863-201803-I/en

ITU Telecommunication Standardization Sector (ITU-T). Recommendation P.910, “Subjective video quality assessment methods for multimedia applications,” April 2008, https://www.itu.int/rec/T-REC-P.910/en.

ITU Telecommunication Standardization Sector (ITU-T). Recommendation J.247, “Objective perceptual multimedia video quality measurement in the presence of a full reference,” August 2008, https://www.itu.int/rec/T-REC-J.247/en.

LLVM Compiler Infrastructure (website), https://llvm.org.

Lassiter, G. Daniel, R. David Slaw, Michael A. Briggs, and Carla R. Scanlan. “The Potential for Bias in Videotaped Confessions,” *Journal of Applied Social Psychology* 22, no. 23 (December 1992):1838–1851. doi:10.1111/j.1559-1816.1992.tb00980.x.

Lederer, Fredric. “The Legality and Practicality of Remote Witness Testimony,” *Practical Litigator,* 22 (September 2009):19–22. http://files.ali-aba.org/thumbs/datastorage/lacidoirep/articles/PLIT0909-Lederer_thumb.pdf.

Pedersens, Philip. TANTBERG H.323, Norway: TANDBERG (March 2010).

Petow, Patricia A. “How to Avoid Video Teleconferencing Hearings,” *Social Security & SSI Disability* (blog), June 26, 2014, http://disabilitydisability.blogspot.com/2014/06/video-teleconferencing-hearings.html.

Quan Huynh-Thu, Marie-Neige Garcia, Filippo Speranza, Philip Corriveau, and Alexander Raake. “Study of Rating Scales for Subjective Quality Assessment of High-Definition Video,” in

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

IEEE Transactions on Broadcasting 57, no. 1 (March 2011):1–14. doi:10.1109/TBC.2010.2086750.

Quikaid (web page). "A video hearing may provide a faster hearing date, but it can be a grave mistake to go this route," accessed 10 September 2018, https://www.quikaid.com/article/vtc-disability-video-hearings.

Russ, John C. *The Image Processing Handbook*. Fourth Edition. (Boca Raton, FL: CRC Press, 2011).

Smith, Christopher J., P.A. (web page). "Video Teleconference Hearing (VTC)," accessed 10 September 2018, https://cjsmithlaw.com/social-security-disability/video-teleconference-hearing-vtc/.

Syfrett, Dykes, & Furr web page. "The Use of Video Teleconferencing for Hearings in Social Security Disability Cases," accessed 10 September 2018, https://www.sdf-law.com/blog/2017/january/the-use-of-video-teleconferencing-for-hearings-i/.

SysTutorials. "tc-netem (8) – Linux Man Pages," accessed 9 November 2017, updated 2017, https://www.systutorials.com/docs/linux/man/8-tc-netem/.

Timony, James P. "Demeanor Credibility," *Catholic University Law Review* 49, no. 4 (Summer 2000):912–913, http://scholarship.law.edu/lawreview/vol49/iss4/2.

Truhelp web page. "Appealing your SSDI claim video teleconference or in person hearing," accessed Sept 10, 2018, https://www.truehelp.com/appealing-your-ssdi-claim-video-teleconference-or-in-person-hearing/.

U.S. Court of Appeals. Case No. 15–1349, Document No. 1613347, filed: 05/16/2016. 70.

United States v. Algere. 457 F. Supp. 2d 695, 700 (E.D. La. 2005).

Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing* 13, no. 4 (April 2004). https://ece.uwaterloo.ca/~z70wang/publications/ssim.pdf.

Wellborn, O.G. III. "Demeanor," *Cornell Law Review*, Vol. 76, No. 5, July 1991, p. 1075.

Wiggins, Elizabeth C. "What We Know and What We Need to Know About the Effects of Courtroom Technology." *William and Mary Bill of Rights Journal* 12, no. 3 (2004): 731-43. http://scholarship.law.wm.edu/cgi/viewcontent.cgi?article=1258&context=wmborj.

Willis, J., and A. Tudorov. "First Impressions–Making Up Your Mind After a 100-Ms Exposure to a Face," *Psychological Science* 17, no. 7 (July 2006):592–598.

APPENDIX G. ACRONYMS AND ABBREVIATIONS

ANT-HIL	Advanced Networking Technologies Lab Hardware in the Loop
AVI	Audio Video Interleave
BSD	Berkley Software Distribution
CLCT	Center for Legal & Court Technology
CLI	Command Line Interface
FEI	Facial Expression Interpretability
GUI	Graphical User Interface
HDMI	High-Definition Multimedia Interface
ITU	International Telecommunications Union
ITU-T	ITU Telecommunication Standardization Sector
JHU/APL	Johns Hopkins University Applied Physics Laboratory
JND	Just Noticeable Difference
LIVE	Laboratory for Image and Video Engineering
LCD	Liquid-Crystal Display
MOS	Mean Opinion Score
MSE	Mean-Square-Error
NetEM	Network Emulation
OpenCV	Open Source Computer Vision Library
OEM	Original Equipment Manufacturer
PC	Personal Computer
PESQ	Perceptual Evaluation of Speech Quality
PEVQ	Perceptual Evaluation of Video Quality
POLQA	Perceptual Objective Listening Quality Analysis
PSNR	Peak Signal-to-Noise Ratio
PSQM	Perceptual Speech Quality Measure
QoE	Quality of Experience
RGB	Red, Green, and Blue
RT&E	Research, Test, and Evaluation
SDI	Serial Digital Interface

SSIM	Structural SIMilarity
VQ	Video Quality
VQEG	Video Quality Experts Group
VTC	Video Teleconferencing
WAN	Wide Area Network