Evaluating Quality and Comprehension of Real-Time Sign Language Video on Mobile Phones

Jessica J. Tran[1], Joy Kim[2], Jaehong Chon[1],
Eve A. Riskin[1], Richard E. Ladner[2], Jacob O. Wobbrock[3]

[1]Electrical Engineering
University of Washington
Seattle, WA 98195 USA
{jjtran, jaehong,
riskin}@ee.washington.edu

[2]Computer Science & Engineering
University of Washington
Seattle, WA 98195 USA
{jojo0808, ladner}@cs.washington.edu

[3]The Information School
DUB Group
University of Washington
Seattle, WA 98195 USA
wobbrock@uw.edu

ABSTRACT
Video and image quality are often objectively measured using peak signal-to-noise ratio (PSNR), but for sign language video, human comprehension is most important. Yet the relationship of human comprehension to PSNR has not been studied. In this survey, we determine how well PSNR matches human comprehension of sign language video. We use very low bitrates (10-60 kbps) and two low spatial resolutions (192×144 and 320×240 pixels) which may be typical of video transmission on mobile phones using 3G networks. In a national online video-based user survey of 103 respondents, we found that respondents preferred the 320×240 spatial resolution transmitted at 20 kbps and higher; this does not match what PSNR results would predict. However, when comparing perceived ease/difficulty of comprehension, we found that responses did correlate well with measured PSNR. This suggests that PSNR may not be suitable for representing subjective video quality, but can be reliable as a measure for comprehensibility of American Sign Language (ASL) video. These findings are applied to our experimental mobile phone application, *MobileASL*, which enables real-time sign language communication for Deaf users at low bandwidths over the U.S. 3G cellular network.

Categories and Subject Descriptors
K.4.2. [**Social Issues**]: Assistive technologies for persons with disabilities; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – Video.

General Terms
Performance, Experimentation, Human Factors.

Keywords
PSNR, video compression, bitrate, spatial resolution, online survey, mobile phones, American Sign Language, Deaf community.

1. INTRODUCTION
Real-time mobile video chat is becoming popular for communication. This enables deaf people to communicate in a language accessible to many of them, American Sign Language (ASL). However, some mobile video chat programs like iPhone's FaceTime [1] only work over Wi-Fi, and other mobile video chat programs like Qik, Fring, Purple, and ZVRS [10,20,21,36] require access to expensive and not widely available 4G cellular networks and smartphones. Also, many cellular networks (AT&T and Verizon) no longer provide unlimited data plans, further limiting access to mobile video calls. To address these limitations, we created an experimental mobile phone application called *MobileASL* [2], which enables Deaf people to communicate in real-time via sign language at low bitrates over the U.S. cellular network. What distinguishes MobileASL is that it is able to transmit over 3G in addition to 4G and Wi-Fi and uses region of interest identification [5] to enable transmission of intelligible sign language video at very low bitrates, making sign language video available to many more devices and people.

Figure 1: One frame of a paired-comparison of 192×144 (left) and 320×240 (right) spatial resolutions transmitted at 10 kbps and displayed at 320×240 pixels.

Research on audiovisual quality [15,34,35] has indicated that when hearing people are shown video with visually detailed scenes at low bitrates, sound becomes increasingly important to compensate. We investigate whether video quality is perceived differently among deaf and non-deaf users since sound cannot be used to compensate for low video quality for deaf[1] people. Since comprehension of video is a subjective measure, objective metrics like peak-signal-to-noise ratio (PSNR), a widely used measure of objective video quality [32], do not necessarily reflect comprehension and subjective quality as perceived by viewers [11,13]. Researchers have tried to create algorithms [17,28,31] to mimic the human visual system to measure subjective quality, but the success at which algorithms reflect users' perceptions varies with users, video content, and data transmission rates. Therefore, we turn to the user to investigate *perception* (between ASL and non-ASL speakers) and *comprehension* (ASL speakers only) of video quality at varying low bitrates and spatial resolutions.

We created and deployed a national video-based online survey to investigate user preferences and comprehension when varying the

ASSETS'11, October 24-26, 2011, Dundee, Scotland, UK.

[1] Using capital "Deaf" is accepted practice when referring to members of Deaf Culture, while lower case "deaf" is used when referring to an individual with hearing loss.

bitrates (10-60 kbps in increments of 10 kbps) and spatial resolutions (192×144 and 320×240) of ASL video that would be transmitted for mobile video phone communication. We seek to answer four questions:

1) When users are shown ASL video encoded at different spatial resolutions and bitrates, which combinations do they prefer?
2) How does the objective video quality measure (PSNR) compare to the subjective video quality preferences for varying bitrates and spatial resolutions?
3) For respondents who are fluent in ASL, does video quality preference influence comprehension of video content with varied spatial resolutions and bitrates?
4) For respondents who are fluent in ASL, how do varied spatial resolutions and bitrates affect their perceived ease/difficulty of comprehension?

In our survey, both ASL and non-ASL speaking respondents overwhelmingly preferred the video quality of the larger spatial resolution at bitrates of 20 kbps ($\chi^2_{1,N=95}$=68.4, p<.0001) and higher. However, the objective PSNR measurements showed a crossover point at 50 kbps and higher, where transmitting the larger spatial resolution (320×240 instead of 192×144) had higher objective video quality than the smaller spatial resolution transmitted at the same bitrates. Despite PSNR not accurately reflecting subjective quality, it did accurately correlate *with comprehension of ASL video*. We found that comprehension was made easier when the larger spatial resolution was transmitted at 50 kbps (Z=100.0, p<.001) and higher, the same crossover point as for the PSNR. These findings and others are presented in detail in our results section.

The main contributions of this paper are identifying that subjective video quality preferences do not differ among ASL and non-ASL speakers; that the perceived ease/difficulty of ASL video comprehension is affected by bitrate and spatial resolution at which video is transmitted; and that PSNR may correlate with perceived ease/difficulty of comprehending ASL video. These results can be used to understand how video comprehension relates to PSNR, which may enable designers of video telephony systems to optimize their choices; for example, to save battery life on mobile devices whose power resources are highly constrained.

2. RELATED WORK

Numerous metrics and algorithms have been created in an attempt to bridge the gap between PSNR and subjective video quality. However, the PSNR has not been shown to accurately represent subjective video quality [8,18,22,30] and a standard subjective metric has not yet been adopted.

Feghali *et al*. [8] created a subjective quality model that takes into account encoding parameters (quantization error and frame rate) and motion speed of video during calculation of their new subjective quality metric. They used the Pearson's correlation, r, which is a measure of how well their subjective model matches subjective video quality, where values closer to 1 indicate a perfect positive linear relation. They were able to achieve, on average (across five videos with different motion levels) an r = .93 when comparing the assessed subjective quality to their new subjective quality metric. For high motion video, such as a football game, the assessed subjective quality compared to the PSNR resulted in r = .57, while the new quality metric resulted in r = .95; however, a smaller difference in r was found for slow motion video. Nemethova *et al*. [18] created a different rule-based algorithm that adapts the PSNR curve to the mean opinion scores (MOS) by scaling, clipping, and smoothing the PSNR results. The new MOS adapted from the PSNR curve was compared to the assessed subjective MOS whose results demonstrated an average r = .89. Both algorithms demonstrated success in increasing the accuracy of measuring subjective video quality; however, both researchers recognize that their algorithms are content-dependent and have higher performance with fast motion video, of which sign language video would be considered one.

Related research by Ciaramello and Hemami [7] developed an objective measure of ASL intelligibility which relies on region-of-interest (ROI) encoding of different areas of video. They encoded ASL video at three different bitrates (20, 45, and 80 kbps) and five ROI settings that vary the allocation of bits to the background and the signer in the foreground during video encoding. This resulted in video with the background appearing blurrier than the ASL signer depending on the bitrate and ROI combinations. In a paired comparison experiment with 12-respondents, they found that at higher bitrates, respondents preferred the background and signer in the foreground to be equal in blurriness; however, at lower encoding bitrates respondents preferred the signer to be less blurry than the background. Our experiment is different than theirs since we are evaluating both subjective video quality *and* comprehension while they only evaluated subjective video quality. We are interested to learn how preferences and comprehension may change with varying spatial resolutions and bitrates since a person may not like a video quality, but still may be able to understand its content.

A related research topic is investigating tolerance of image artifacts when lowering bitrates and image resolutions. Bae *et al*. [4] conducted a 7-respondent experiment that assessed absolute perceived quality and relative perceived quality of compressed images at different bitrates. In the absolute perceived quality assessment, respondents were shown uncompressed images and asked to score the image on a 5-point Likert scale. Next, compressed sets of images were presented to the participant, who selected the one image that they preferred the most. Bae *et al*. discovered that as bitrates decrease, respondents preferred to maintain image quality by selecting a lower image resolution. Respondents were willing to accept an increase in image distortion (compression noise) introduced by the coding algorithms when shown an image at smaller spatial resolutions.

A similar research topic has been conducted to understand how varying frame rate and display size of ASL video affects comprehension when shown on a computer. Hooper *et al*. [12] conducted a subjective study to determine if varying frame rate and display size of ASL video would impact learning comprehension. Their study investigated three frame rates (6, 12, and 18 fps) and three video display sizes (240×180, 320×240, and 480×360) with the bitrate for each video held constant at 700 kbps. They found that the display size of video did not affect comprehension, but varying the frame rates did. Our study is different than Hooper *et al.*'s because we are interested in comprehension of video at bitrates ten times less than what they used in their study and transmitting smaller spatial resolutions at a constant frame rate. Our previous research on MobileASL [5] has investigated varying frame rates [6,26] for data transmission and will not be elaborated on. We expand by varying spatial resolutions and bitrates to investigate subjective video quality preferences and comprehension all while comparing these results to PSNR measurements.

3. PSNR CURVES

Selecting a specific spatial resolution and bitrate combination to transmit video on MobileASL is important because there are tradeoffs with computational complexity, video quality, and resource availability on a cell phone such as battery life and data rate consumption. Larger video resolutions and higher bitrates result in higher video quality at the expense of increased computational power to transmit the data in real time. Before we can investigate how resource allocation is affected by video transmission, we need to determine at which bitrates and spatial resolutions we can get high enough video quality for intelligible conversations.

Despite the fact that PSNR may not be suitable for measuring subjective video quality, it still is a reasonable measurement of video quality when used across the same content [25]. We calculated PSNR of two different spatial resolutions (192×140 and 320×240 pixels) and 15 bitrates (10-150 kbps in increments of 10 kbps) of the same ASL video. The smaller spatial resolution was transmitted at 192×140 pixels and then enlarged and displayed at 320×240 pixels using bilinear interpolation [27] before PSNR was calculated.

The same 12-second video clip of a local deaf woman signing at her natural signing pace with a stationary background was used in the calculation of PSNR. The video was recorded at 320×240 pixels at 15 fps. Duplicate videos were created at the smaller spatial resolution before calculating the PSNR. The *x*264 codec, an open source version of H.264 codec, was used to compress the videos at each spatial resolution and bitrate combination [3,24]. As Figure 2 demonstrates, the PSNR values for each spatial resolution increase monotonically with increasing bitrate.

We found that the PSNR curves demonstrated a crossover point where, at lower bitrates (40 kbps and below), the smaller spatial resolutions had higher PSNR values than the larger spatial resolution. Visual inspection of the same ASL video (displayed at the same size) transmitted at lower bitrates (10-40 kbps) showed more blocky artifacts in videos sent at 320×240 pixels than at 192×144. The crossover in the PSNR plots occurred because at very low bitrates, the higher resolution video is quantized more heavily and thus has very poor visual quality (such as blockiness and loss of fine details). The same videos at lower spatial resolutions are not quantized as heavily which results in higher measured video quality. As bitrates increase, the higher resolution has higher measured video quality than the smaller spatial resolutions. This is due to blurriness from enlarging the video. The crossover of PSNR curves has been found in other video compression techniques [16,19,29], but the results, to our knowledge, have not been used to evaluate human comprehension, which, along with subjective quality measures, is the focus of our online survey.

4. ONLINE SURVEY METHOD

From a technological perspective, transmitting video at the smaller spatial resolution and at the lowest bitrates takes the least amount of computational power and resources; however, without feedback from users, we cannot confirm that sign language communication with this video is intelligible.

We created and deployed a national three-part online survey to investigate user preferences and comprehension when varying the bitrates (10-60 kbps in increments of 10 kbps) and spatial resolutions (192×144 and 320×240) of ASL video. We did not consider bitrates higher than 60 kbps since the larger spatial resolution always had higher video quality than the smaller spatial resolution.

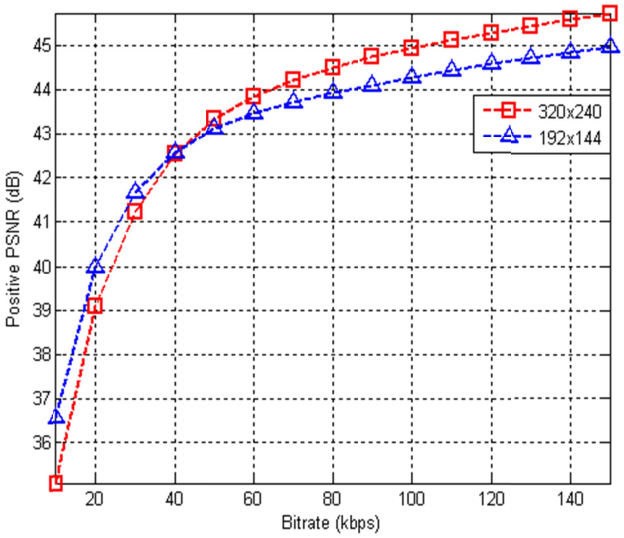


Figure 2: PSNR(dB) *vs.* Bitrate (kbps) for spatial resolutions displayed at 320×240 pixels. Higher PSNR means higher objective video quality. Whether it means higher subjective perception of quality is a topic of this research.

The online survey began by asking participants to self-report their fluency in ASL. The survey asked different questions depending on the response to this question. Part 1 was a paired-comparison experiment which investigated the subjective video quality preferences of ASL speakers and non-ASL speakers (see Figure 3). Part 2 was a single-stimulus experiment which examined comprehension of ASL video of varying bitrates and spatial resolutions (ASL speakers only) (see Figures 4 and 5). Finally, part 3 asked demographic questions.

To determine how subjective video quality preference differs between ASL speakers and non-ASL speakers, it was important to get an equal number of ASL and non-ASL speaking respondents. We selected an online survey over a laboratory study because an online survey is accessible to most people with Internet access, so more respondents could be included from across the nation.

4.1 Videos Used in Online Survey

4.1.1 Videos in Part 1

The same 12-second video clips used to measure PSNR (see section 3, above) of ASL video were used in part 1 of the survey. A 12-second video duration was used because it was long enough for respondents to make a video preference selection while keeping the overall survey manageable to complete in 4-7 minutes. Recall that all videos were transmitted at their respective spatial resolution (192×144 and 320×240) at varied bitrates, and then displayed at 320×240 pixels (with the smaller spatial resolution enlarged using bilinear interpolation).

4.1.2 Videos in Part 2

Twelve different video clips of the same local deaf woman signing different short stories at her natural signing pace were used.

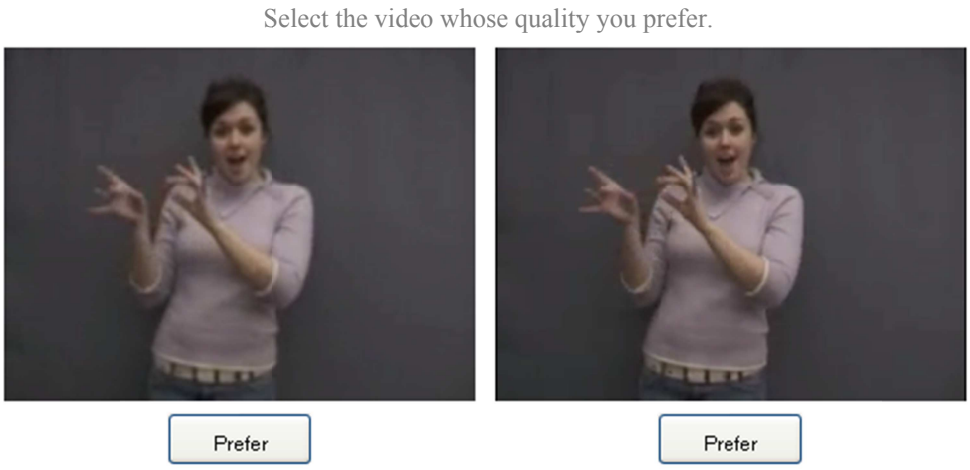


Figure 3: Screenshot of one 12-second video pair from the paired-comparison experiment. Respondents selected which video they preferred to watch.

All videos were recorded with the same parameters listed in section 3. Each video was again truncated to the first 12-seconds of the story to keep the overall duration of the survey manageable and to test respondents with comprehension question about that segment. A duplicate set of the twelve videos were created and downsampled to a spatial resolution of 192×144 pixels. Next, the *x*264 codec was used to compress the videos at the six different bitrates [3,24].

4.2 Paired-Comparison Experiment

As Figure 3 demonstrates, part 1 of the survey used a paired-comparison method with simultaneous presentation as described in prior work [14]. For each of the six bitrates, a pair of videos (each at the two different spatial resolutions) was shown. This yields six pair-wise combinations, one at each bitrate. The videos were shown side-by-side on the same screen with synchronous playback. Respondents could watch the video pairs repeatedly until a selection was made. Each of the six pairs was presented twice, switching the left/right display order to counterbalance and prevent bias from video placement. None of the test pairs contain videos at different bitrates, since previous research [5] confirmed that higher bitrates were always selected when given the option. This study design resulted in twelve trials per participant. Randomization was done with an algorithm that randomly selected the next video after eliminating the previous selection. During each trial, respondents were asked to select the video whose quality they preferred. To make sure respondents watched the video pairs, they could not select a preferred video until four seconds after a video pair began playing. In addition to recording which video the participant preferred, we also recorded the amount of time it took for a participant to select his or her choice.

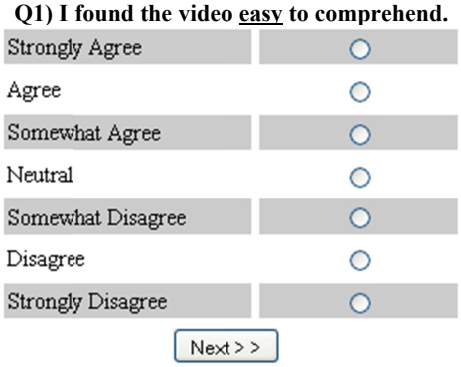


Figure 4: Q1 was a 7-point Likert scale for the ease of comprehension. Q1 was shown after the video was removed from the screen.

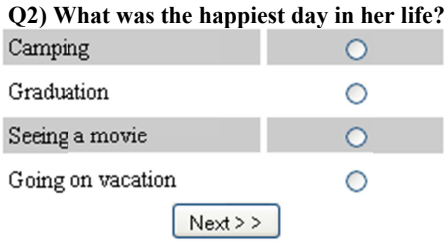


Figure 5: Q2 asked a simple comprehension question pertaining to the video shown. Q2 was shown after Q1 was removed from the screen.

4.3 Single Stimulus Experiment

A single stimulus experiment, as described in prior work [14], was used to evaluate comprehension of ASL video transmitted and encoded at each combination of spatial resolution and bitrate. These combinations yield twelve videos in the single stimulus experiment. Before beginning part 2, they were shown a practice video to familiarize themselves with the layout.

Each video was shown once (without the option to repeat the video), then removed from the screen and replaced by two questions shown one at a time. Figure 4 is an example of question 1 which asked the participant to rate their agreement/disagreement on a 7-point Likert scale with the statement, "I found the video easy to comprehend." The 7-point Likert scale was shown in descending vertical order from *strongly agree* to *strongly disagree*. The word 'difficult' replaced the word 'easy' for every other respondent, but always remained the same within a respondent. This approach prevented bias from respondents' interpretations of "easy" or "difficult." Figure 5 is an example of question 2 which asked a trivial comprehension question pertaining to the video shown. Since the ease/difficulty of comprehension varied with each 12-second video segment, the

comprehension questions were only used as a way to confirm that the participant had been paying attention to the video.

4.4 Demographic Questions

After respondents completed parts 1 and 2, they were asked background questions which included: "What is your age?"; "What is your gender?"; "Do you own a cell phone or Blackberry?"; "Do you text message on the cell phone or Blackberry?"; "If applicable, what operating system is on your cell phone?"; "Do you video chat?"; "If applicable, which video chat program do you use?"

ASL speaking respondents were also asked: "If applicable, how many years have you spoken ASL?"; "If applicable, from whom did you learn ASL?"; "What language do you prefer to communicate with family?"; "What language do you prefer to communicate with friends?"; "Are you Deaf?"; "Do you use a video phone?"; "Do you use video relay services?"

5. RESULTS

Recall, at the start of the survey, respondents self-declared their fluency in ASL. In part 1 of the survey, we investigated (1) the preferences of both ASL speakers and non-ASL speakers for spatial resolution as bitrates varied, and (2) how subjective video quality preferences compared to measured PSNR values. In part 2 of our survey, we were interested in whether comprehension of ASL video content by respondents fluent in ASL was affected by transmission bitrate and spatial resolution.

A total of 103 respondents completed the survey; however, in part 1, we eliminated results from those who used internet browsers incompatible with our survey. We kept results from respondents who completed part 1 but failed to finish the entire survey (part 2 and demographics sections). In part 1, we analyzed data from 95 respondents: 56 ASL speakers (30 men, 15 women, and 11 who did not specify) and 39 non-ASL speakers (13 men, 25 women, and 1 who did not specify). Their age ranged from 18-71 years old (mean: 37 years). Of the respondents who self-reported fluency in ASL, 41 were deaf, 35 self-declared using ASL as their daily language, and the number of years they have spoken ASL ranged from 3-58 years (mean: 26 years). Seventy-eight respondents (43 ASL, 35 non-ASL) owned a cell phone, and 72 of those cell phone owners (43 ASL, 29 non-ASL) used it to text message.

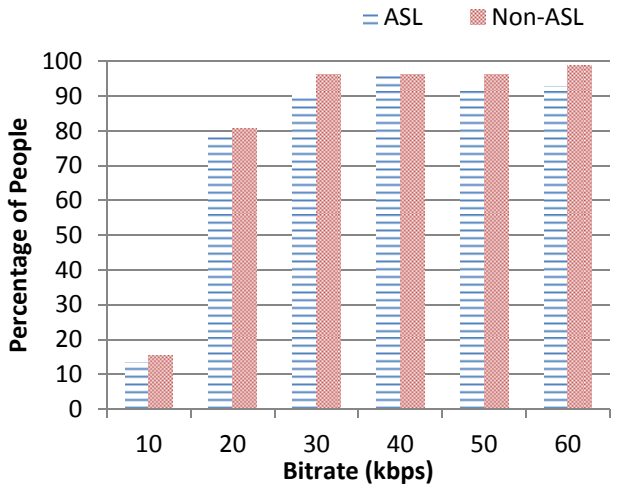


Figure 6: Percentage of People *vs.* Bitrate (kbps) who selected 320×240 instead of 192×144 spatial resolution in the paired-comparison experiment. Data is from 56 ASL speakers and 39 Non-ASL speakers.

For part 2 of the survey, we analyzed data from 53 respondents (33 men, 18 women, and 2 who did not specify). Their age ranged from 18-71 years old (mean: 27 years) and all but five respondents were deaf. The self-reported number of years they have spoken ASL ranged from 3-58 years (mean: 27 years). Forty-one respondents indicated they use ASL as their daily language. Finally, 48 respondents indicated they own a cell phone, with all of them using text messaging, and all but three respondents said they use video phones and/or video relay services.

5.1 Subjective Video Quality Preferences

Respondents were asked to select which video they preferred when presented with two videos playing simultaneously side-by-side at the same bitrates. Figure 6 shows the percentage of people *vs.* bitrate who selected the 320×240 spatial resolution over the 192×144 spatial resolution by ASL and non-ASL speaking respondents.

A one-sample Chi-Square test was performed to test whether the proportion of subjects who picked the 320×240 spatial resolution *vs.* the 192×144 spatial resolution was significantly different than chance at each bitrate (10-60 kbps in increments of 10 kbps). Recall that both videos were *displayed* at the same spatial resolution (320×240).

At 10 kbps, both subject groups overwhelmingly preferred the video quality of the lower 192×144 spatial resolution over the 320×240 spatial resolution ($\chi^2_{1,N=95}$=97.347, p<.0001). At transmission bitrates of 20 kbps and higher, both subject groups preferred the video quality of the 320×240 spatial resolution ($\chi^2_{1,N=95}$=68.40, p<.0001).

5.2 Video Comprehension

Respondents were asked to rate their perceived ease/difficulty of comprehending each of the twelve videos on a 7-point Likert scale. Recall that the wording of this question alternated *between* respondents, but remained the same *within* each participant.

Nonparametric analyses were used to analyze our 7-point Likert scale responses for rating the perceived ease/difficulty of comprehension. Since we gathered ordinal and dichotomous response data, a Friedman test [9] was used to analyze the main effect of bitrate and spatial resolution on comprehension. Separate Wilcoxon tests [33] were performed to investigate the effect of spatial resolution *within* each bitrate.

The Friedman test indicated a significant main effect of spatial resolution on video comprehension ($\chi^2_{1,N=53}$=8.33, p<.01). The Friedman test also indicated a significant main effect of bitrate on video comprehension ($\chi^2_{5,N=53}$=146.15, p<.0001).

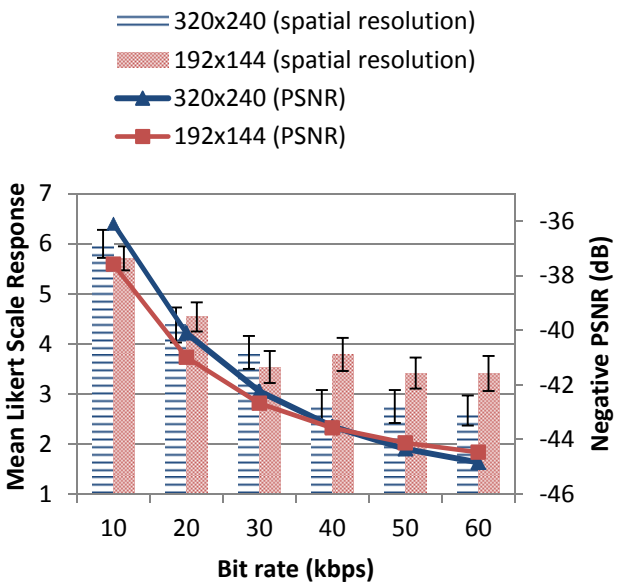
Wilcoxon tests with Bonferroni correction were performed *within* each bitrate to identify the effect of spatial resolution on comprehension. Of the 53 respondents, 24 were asked to rate the difficulty of comprehension and 29 were asked to rate the ease of comprehension. The results of the Wilcoxon test for the perceived ease/difficulty of comprehension are presented separately, below.

Table 1: Mean Likert Scale responses (1-7) for difficulty of comprehending video quality. Note *lower* Likert scores correspond to *less* perceived difficulty.

	Spatial Resolution			
	320×240		192×144	
Bitrate	Mean	Std. Error	Mean	Std. Error
10	6.00	0.28	**5.71**	0.24
20	**4.38**	0.35	4.54	0.29
30	3.83	0.33	**3.54**	0.32
40	**2.75**	0.33	3.79	0.33
50	**2.75**	0.33	3.42	0.31
60	**2.67**	0.30	3.41	0.35

Table 2: Mean Likert Scale responses (1-7) for ease of comprehending video quality. Note *higher* Likert scores correspond to *easier* perceived comprehension.

	Spatial Resolution			
	320×240		192×144	
Bitrate	Mean	Std. Error	Mean	Std. Error
10	2.90	0.31	**3.55**	0.28
20	**5.10**	0.29	4.72	0.29
30	5.34	0.26	**5.48**	0.26
40	**5.90**	0.25	5.41	0.23
50	**6.27**	0.19	5.48	0.22
60	**6.34**	0.14	5.62	0.20



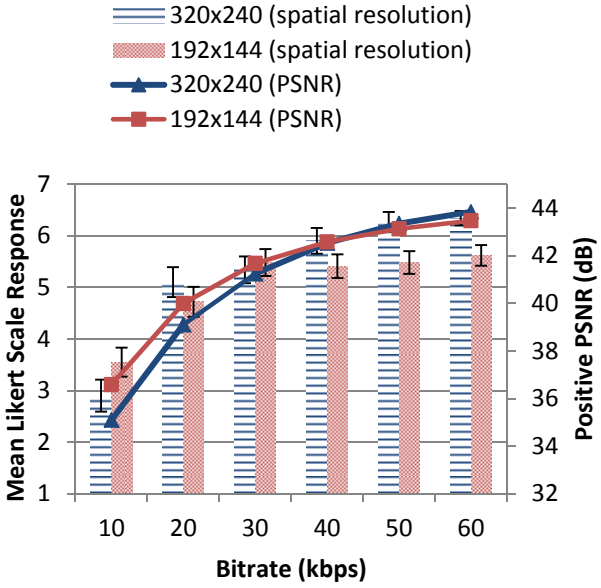


Figure 7: Double *y*-axis plot of 7-point Likert scale. Negative PSNR values of spatial resolutions and bitrates. Lower Likert scores correspond to less difficulty and lower PSNR values correspond to higher video quality. Notice a negative PSNR crossover point occurs at 40 kbps.

Figure 8: Double *y*-axis plot of 7-point Likert scale. Positive PSNR values of spatial resolution and bitrate. Higher Likert scores correspond to more ease and higher PSNR values correspond to higher video quality. Notice a positive PSNR crossover point occurs at 40 kbps.

5.2.1 Rating Difficulty of Comprehension

Recall that about half of the respondents saw a 7-point Likert scale concerning the *difficulty* of comprehension, ranging from 1 (strongly disagree), *i.e.*, less difficult to comprehend, to 7 (strongly agree), *i.e.*, more difficult to comprehend. Table 1 shows the mean Likert scale response for the difficulty of comprehending the ASL video transmitted at each bitrate and spatial resolution and displayed at 320×240 pixels.

Figure 7 is a double *y*-axis plot of the mean Likert responses and the negative PSNR values for each bitrate and spatial resolution. Notice that the PSNR values are *negative*, where *lower* values correspond to *higher video quality*.

Comprehension was significantly less difficult at 60 kbps for the 320×240 spatial resolution than the 192×144 spatial resolution ($Z=35.0$, $p<.01$). However, changing the spatial resolution within other bitrates did not indicate more difficulty in comprehension. For example, Table 1 and Figure 7 indicated a large difference of mean Likert scores at 40 kbps, but changing the spatial resolution within that bitrate was not significant in affecting the difficulty of comprehension ($Z=48.5$, *n.s.*).

5.2.2 Rating Ease of Comprehension

Recall that about half the respondents saw a 7-point Likert scale concerning the *ease* of comprehension, ranging from 1 (strongly disagree), *i.e.*, less easy to comprehend, to 7 (strongly agree), *i.e.*, more easy to comprehend. Table 2 shows the mean Likert scale response for the ease of comprehending ASL video transmitted at each bitrate and spatial resolution and displayed at 320×240 pixels.

Figure 8 is a double *y*-axis plot of the mean Likert responses and the positive PSNR values for each bitrate and spatial resolution. Notice that the PSNR values are *positive*, where *higher* values correspond to *higher video quality*.

Transmitting at 320×240 spatial resolution rather than at a 192×144 spatial resolution at 50 and 60 kbps was significantly easier to comprehend ($Z=100.0$, $p<.001$ and $Z=88.5$, $p<.001$, respectively). This result is also shown in the PSNR curve in Figure 8; at 50 kbps and 60 kbps, the positive PSNR values were higher for the larger spatial resolution. However, changing the spatial resolution within other bitrates did not make the content easier to understand. Even though Table 2 and Figure 8 indicate a large difference of mean Likert score at 10 kbps, changing the spatial resolution within that bitrate was not significant in affecting comprehension ($Z=45.5$, *n.s.*).

6. DISCUSSION

We compared the video preference results from part 1 to PSNR measurements, which reinforced that PSNR may not accurately reflect subjective video quality. The PSNR values suggested that bitrates at 40 kbps and lower spatial resolution of 192×144 pixels had higher quality than the 320×240 spatial resolution; however, subjective user preferences revealed that at 20 kbps and higher, the larger spatial resolution was preferred. This finding is not unexpected since PSNR does not account for compression artifacts (blockiness and Gibbs's phenomena [23]). Also, visual inspection of each pair of videos showed that at bitrates 20 kbps and higher, enlarging the smaller spatial resolution to display at 320×240 pixels caused the video to appear more blurry than when simply transmitting the larger spatial resolution.

One might expect that the same bitrates and spatial resolutions indicated as preferred in part 1 would similarly influence comprehension of content; that is, that respondents would indicate greater ease (or less difficulty) of comprehension when shown video at a 320×240 spatial resolution at bitrates of 20 kbps and higher. However, transmitting either spatial resolution sent at 10-50 kbps had no effect on making comprehension more difficult. At 60 kbps only, respondents expressed that transmitting the larger spatial resolution made the content significantly less difficult to comprehend. This result was the same among the respondents who were asked to rate the ease (rather than the difficulty) of comprehension. Neither of the two spatial resolutions, at bitrates of 10 to 40 kbps, made comprehending the video easier. However, at 50 and 60 kbps, respondents did indicate that transmitting the larger spatial resolution made comprehension easier. When comparing these findings to the PSNR curves (Figures 8 and 9), we see that PSNR measurements may accurately reflect the perceived ease/difficulty at which respondents rated comprehension of ASL video. The PSNR curves showed a threshold where at 50 kbps and higher, transmitting the larger spatial resolution produces better video quality than transmitting and enlarging the smaller spatial resolution. The results of the survey agree with this and also indicate that at 50 kbps and higher, video comprehension was made easier.

These results suggesting that PSNR may be a reliable measure for comprehensibility of ASL video may be valuable in selecting the spatial resolution and bitrate for mobile video telephony. Having knowledge of how PSNR relates to comprehension, especially for sign language video, can influence how video is transmitted on mobile phones using 3G networks. When possible, selecting the smaller spatial resolution at the PSNR crossover point provides intelligible video while keeping computational complexity and cost of video transmission low. For MobileASL, transmitting video at 40 kbps at 192×144 spatial resolution would be sufficient to hold an intelligible conversation while saving limited computing resources.

7. CONCLUSION AND FUTURE WORK

In this work, we investigated how varying bitrates and spatial resolutions of ASL video affect subjective video quality (for both self-reported ASL speakers and non-ASL speakers) and comprehension of video content (ASL speakers only). We found that our respondents' preferences for video spatial resolutions at different bitrates did not agree with the results of the calculated PSNR values of measured video quality. Whether or not respondents were fluent in ASL did not impact their preference for bitrate and spatial resolution; both groups selected the 320×240 spatial resolution over the 192×144 spatial resolution at 20 kbps and higher. However, we did find a main effect where changing the spatial resolution and bitrate significantly impacted perceived ease/difficulty of comprehension. Closer inspection of which spatial resolution and bitrates significantly impacted comprehension revealed that the 320×240 spatial resolution sent at 50-60 kbps improved the ease of comprehension. A notable finding was that PSNR may correlate with rating the perceived ease/difficulty of comprehension at higher bitrates and spatial resolutions. Therefore, the recommendation for MobileASL is to transmit video at 192×144 spatial resolution at 40 kbps to provide intelligible sign language video while keeping computational costs low.

For future work, we would like to see how our findings can be applied to improve consumption of mobile phone resources such as battery life and data consumption of metered cell phone plans. We are particularly interested to learn if behavioral changes occur when users are aware of how they consume resources and, if given the option, would users elect to lower bitrates and spatial resolution to gain more battery life or conversation time.

9. ACKNOWLEDGEMENTS

Thanks to Anna Cavender, Jessica Belwood, Frank Ciaramello, Katie O'Leary, Sorenson VRS, and our respondents. This work was supported by the National Science Foundation under grant IIS-0811884.

10. REFERENCES

[1] Apple FaceTime. Retrieved 4 29, 2011, from Apple FaceTime: http://www.apple.com/mac/facetime/

[2] *MobileASL. University of Washington*. (2011). Retrieved from http://mobileasl.cs.washington.edu/

[3] Aimar, L., Merritt, L., Petit, E., Chem, M., Clay, J., Rullgrd, M., et al. (2005). x264 - a free h264/AVC encoder.

[4] Bae, S., Pappas, T., & Juang, B. (2006). Spatial Resolution and Quantization Noise Tradeoffs for Scalable Image Compression. *IEEE International Conference Acoustics, Speech, and Signal Processing.*

[5] Cavender, A., Ladner, R., & Riskin, E. (2006). MobileASL: Intelligibility of Sign Language Video as Constrained by Mobile Phone Technology. *Proceedings of ASSETS 2006: The Eighth International ACM SIGACCESS Conference on Computers and Accessibility.* Portland, OR.

[6] Cherniavsky, N., Chon, J., Wobbrock, J., Ladner, R., & Riskin, E. (2007). Variable Frame Rate for Low Power Mobile Sign Language Communication. *Proceedings of ASSETS 2007: The Ninth International ACM SIGACCESS*

Conference on Computers and Accessibility, (pp. 163-170). Tempe, AZ.

[7] Ciaramello, F. & Hemami, S. (2011, January). Quality versus Intelligibility: Studying Human Preferences for American Sign Language Video. *Proceedings in SPIE Volume 7865, Human Vision and Electronic Imaging.*

[8] Feghali, R., Speranza, F., Wang, D., & Vincent, A. (2007, March). Video Quality Metric for Bit Rate Control via Joint Adjustment of Quantization and Frame Rate. *53*(IEEE Transactions on Broadcasting).

[9] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association 32* (200), 675-701

[10] *Fring*. Retrieved 4 29, 2011, from http://www.fring.com/

[11] Girod, B. (1993). What's wrong with mean-squared error? *Digital images and human vision*, 207-220.

[12] Hooper, S., Miller, C., Rose, S., & Veletsianos, G. (2007). The Effects of Digital Video Quality on Learner Comprehension in an American Sign Language Assessment Environment. *Sign Language Studies, 8*(Sign Language Studies), 42-58.

[13] Huynh-Thu, Q., & Ghanbari, M. (2008). Scope of validity of PSNR in image/video quality assessment. The Institution of Engineering Technology.

[14] ITU. (September 1999). p.910: Subjective video quality assessment methods for multimedia applications.

[15] Jumiski-Pyykko. (2005). Evaluation of Subjective Video Quality of Mobile Devices. *Multimedia Proceedings of the 13th annual ACM international conference on Multimedia* . Singapore.

[16] Lin, W., & Dong, L. (2006, September). Adaptive Downsampling to Improve Image Compression at Low Bit Rates. *IEEE Transcations on Image Processing. 15.*

[17] Masry, M., & Hemami, S. (2003, January). CVQE: A metric for continuous video quality evaluation at low rates. *Proceedings in SPIE: Human Vision and Electronic Imaging*.

[18] Nemethova, A., Ries, M., Zavodsky, M., & Rupp, M. (2006). PSNR-Based Estimation of Subjective Time-Variant Video Quality for Mobiles. *Measurement of Audio and Video Quality in Networks.*

[19] Nguyen, V., Tan, Y., Z, & Lin, W. (2006). Adaptive Downsampling/Upsamping for better video compression at low bit rate. *IEEE International Symposium on Circuits and Systems.*

[20] *Purple VRS on Your Devices.* (Purple Communications.) Retrieved 7 31, 2011, from http://www.purple.us/

[21] *Qik*. (Qik, Inc.) Retrieved 4 29, 2011, from http://qik.com/

[22] Reiter, U., & Korhonen, J. (2009). Comparing Apples and Oranges: Subjective Quality Assessment of Streamed Video with Different Types of Distortion. (IEEE).

[23] Radaelli-Sanchez, Baraniuk, R. (2010). Gibbs's Phenomena. http://cns.org/content/m10092/latest

[24] Richardson, I. (2004). vocdex: H.264 tutorial white papers.

[25] Thu, H., & Ghanbari, M. (2008). Scope of Validity of PSNR in image/video quality assessment. *44*.

[26] Tran, J. J., Johnson, T. W., Kim, J., Rodriguez, R., Yin, S., Riskin, E., Ladner, R., Wobbrock. J., (2010). A Web-Based User Survey for Evaluating Power Saving Strategies for Deaf Users of MobileASL. *Proceedings of ASSETS 2010: The 12th International ACM SIGACCESS Conference on Computers and Accessibility.* Orlando, FL.

[27] Vision Sytems Design. Retrieved 5 03, 2011. Understanding image-interpolation techniques.

[28] VQEG. (2003, September). Final report from the video quality experts group on the validation of objective models of video.

[29] Wang, R. Chien, M., Chang, P. Adaptive Down-Samping Video Coding. VQEG. (2010). *Proceedings in SPIE 7542.*

[30] Wang, Z., Bovik, A., & Lu, L. (2002). Why is Image Quality Assessment so Difficult? *IEEE Acoustic, Speech, and Signal Processing.*

[31] Wang, Z., Lu, L., & Bovik, A. (2004, February). Video quality assessment based on structural distortion measurement. *19*(Signal Processing: Image Communication special issue on Objective video quality metrics), 121-132.

[32] Wiegang, T., Schwarz, H., Joch, A., Kossentini, F., & Sullivan, G. (2003). Rate-constrained coder control and comparison of video coding standards. *IEEE Transactions Circuits Systems Video Technology, 13*(7), 688-703.

[33] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin 1* (6), 80-83.

[34] Winkler, S., & Faller, C. (2005). Audiovisual Quality Evaluation of Low-Bitrate Video. *SPIE/IS&T Human Vision and Electronic Imaging*, *5666*, pp. 139-148. San Jose.

[35] Winkler, S., & Faller, C. (2005). Maximizing Audiovisual Quality at Low Bitrates. *Workshop on Video Processing an Quality Metrics for Consumer Electronics.* Scottsdale, AZ.

[36] *ZVRS*. (ZVRS Communication Service for the Deaf, Inc.) Retrieved 7 31, 2011, from http://www.zvrs.com/z-series/